Psychological Review
1992, Vol. 99, No. 3, 418–439

0033-295X/92/$3.00

Propositional Reasoning by Model

P. N. Johnson-Laird
Princeton University

Ruth M. J. Byrne
Department of Computer Science
University College, Dublin
Dublin, Ireland

Walter Schaeken
University of Leuven
Leuven, Belgium

This article describes a new theory of propositional reasoning, that is, deductions depending on *if, or, and,* and *not.* The theory proposes that reasoning is a semantic process based on mental models. It assumes that people are able to maintain models of only a limited number of alternative states of affairs, and they accordingly use models representing as much information as possible in an implicit way. They represent a disjunctive proposition, such as "There is a circle or there is a triangle," by imagining initially 2 alternative possibilities: one in which there is a circle and the other in which there is a triangle. This representation can, if necessary, be fleshed out to yield an explicit representation of an exclusive or an inclusive disjunction. The theory elucidates all the robust phenomena of propositional reasoning. It also makes several novel predictions, which were corroborated by the results of 4 experiments.

Propositional reasoning is ubiquitous in daily life. It consists of combining information from propositions containing such connectives as *if, and, or,* and *not.* For example, take the following premises:

If there is fog, then the plane will be diverted.

There is a fog.

It is easy to draw the following conclusion:

The plane will be diverted

It is natural to suppose that the mind must contain a corresponding formal rule of inference:

If A, then B.

A

∴ B

It is also natural to suppose that the inference proceeds by matching the logical form of the premises to this rule of *modus ponens.* The rule can then be used to derive the conclusion.

The dominant theoretical tradition is indeed that human beings are equipped with formal rules of inference that enable them to make deductions. Versions of such theories have been proposed by most of the those who have worked on the psychology of deductive reasoning. The idea goes back to Boole in the nineteenth century, who wrote of his formal calculus for propositional reasoning, "The laws we have to examine are the laws of one of the most important mental faculties. The mathematics we have to construct are the mathematics of the human intellect" (1847/1948, p. 7). In our era, the formal view has been advocated by Piaget and his colleagues: "Reasoning is nothing more than the propositional calculus itself" (Inhelder & Piaget, 1958, p. 305). Piaget's views about logic were idiosyncratic (e.g., see Braine & Rumain, 1983), but more recent proponents of formal rules have based their systems on orthodox logic. In particular, they have used the logical method of natural deduction, which has separate rules of inference for each of the connectives, *not, if, and,* and *or.* Many theorists have proposed such accounts of the psychology of propositional reasoning (e.g., Braine, 1978; Braine, Reiser, & Rumain, 1984; Johnson-Laird, 1975; Macnamara, 1986; Osherson, 1974–1976, 1975; Pollock, 1989; Rips, 1983, 1988; Sperber & Wilson, 1986). They all hold the view aptly expressed by Rips (1983) in the following terms:

> . . . deductive reasoning consists in the application of mental inference rules to the premises and conclusion of an argument. The sequence of applied rules forms a mental proof or derivation of the conclusion from the premises, where these implicit proofs are analogous to the explicit proofs of elementary logic (p. 40).

Henceforth, we will refer to these accounts of reasoning as *rule theories.*

Our view of deductive competence is that people are rational in principle, but they err in practice. Any set of deductive premises yields an infinite number of valid conclusions, but most of them are banal, such as an arbitrary number of conjunctions of

Walter Schaeken is a research assistant funded by the Belgian National Fund for Scientific Research.

We thank Jonathan Evans for his advice on the phenomena of propositional inference and Malcolm Bauer for carrying out the multiple regression. We are also grateful to Martin Braine, who abides strictly by Marquis of Queensbury rules. Finally, we thank Mark Keane, Steve Palmer, Robert Sternberg, and anonymous referees for their criticisms of an earlier version of this article.

Correspondence concerning this article should be addressed to P. N. Johnson-Laird, Department of Psychology, Princeton University, Princeton, New Jersey 08544.

a premise with itself. Logically untutored individuals never draw such conclusions. In general, they eschew conclusions that contain less semantic information than premises. Hence, suppose they are asked what follows logically from the following premise:

Anne is at the party and Alan is at the game.

They do not spontaneously draw the conclusion:

Anne is at the party.

Similarly, they seek conclusions that are more parsimonious than the premises. Hence, suppose they are asked what follows from the following premises:

Betty is here.

Brian is at work.

They do not spontaneously draw the conclusion:

Betty is here and Brian is at work.

They also do not bother to repeat in their conclusions what is asserted categorically by a premise (cf. Grice, 1975). They try instead to draw conclusions that make explicit some information only implicit in the premises. In short, to deduce is to maintain semantic information, to simplify, and to reach a new conclusion. Where there is no valid conclusion that meets these three constraints, logically untrained individuals declare that nothing follows from the premises (Johnson-Laird, 1983). Any theory of how people reason should accordingly reflect these constraints, though they be may emergent properties of other principles.

In this article, we aim to present a new explanation of propositional reasoning. Our hypothesis is that the underlying deductive machinery depends not on syntactic processes that use formal rules but on semantic procedures that manipulate mental models. This theory is in part inspired by the model-theoretic approach to logic. Semantic procedures construct models of the premises, formulate parsimonious conclusions from them, and test their validity by ensuring that no alternative models of the premises refute them. This approach is akin to the analysis of problem solving as a heuristic search through a problem space in which each state corresponds to a mental model (Newell, 1990; Newell & Simon, 1972; Simon, 1990).

Various sorts of mental models have been postulated as underlying deduction (e.g., see Erickson, 1974; Guyote & Sternberg, 1981; Levesque, 1986; Newell, 1981; Polk & Newell, 1988). Our view is that models have a structure that corresponds directly to the structure of situations. Each individual in a situation is represented by a corresponding mental token, and the properties of individuals and the relations among them are likewise modeled in an isomorphic way (see Johnson-Laird, 1983, p. 419–447). This theory has been applied to spatial reasoning (Byrne & Johnson-Laird, 1989), to reasoning with single quantifiers (see Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1989), and to reasoning with multiple quantifiers (Johnson-Laird, Byrne, & Tabossi, 1989). It has not, however, been applied to propositional reasoning. Hence, as many critics have pointed out (Braine et al., 1984; Evans, 1987; Rips, 1986, 1990), the model theory has been radically incomplete. In the present article, which supersedes the suggestions in Johnson-Laird and Byrne (1991), we remedy this deficiency. We present a comprehensive model theory of propositional reasoning.

Our plan is to consider those connectives that can be accommodated within rule theories, namely, *not, and, or,* and *if.* We begin by outlining the distinction in logic between rules and models. We describe some representative rule theories in psychology and the contrasting model theory, including its computer implementation. This theory motivates a reanalysis of the major experimental studies, and we show how it explains their principal phenomena. The theory leads to novel predictions, and we report some experiments designed to test them. Finally, we consider the chief differences between rules and models.

Rules and Models in Logic

Logicians distinguish between reasoning based on formal rules of inference (proof-theoretic methods) and reasoning based on models (model-theoretic methods). A proof-theoretic method uses formal rules of inference to derive conclusions from premises in a syntactic way. Here, for example, is a formal rule of inference for inclusive disjunction:

A or B, or both.

Not-A.

Therefore, B,

where A and B can be any propositions. The rule can be used to make the following deduction:

Lisa is in Cambridge or Ben is in Dublin, or both.

Lisa is not in Cambridge.

Therefore, Ben is in Dublin.

The disjunctive rule is part of most psychological theories based on formal rules (e.g., Braine, 1978; Johnson-Laird, 1975; Rips, 1983).

A formal calculus can be given a semantic interpretation in terms of models, and the standard model-theoretic method for the propositional calculus is based on truth tables. The meaning of each connective is specified by a truth table. An inclusive disjunction of two propositions, A or B or both, is true provided that at least one of the two propositions is true, and is false only if they are both false. This truth-functional definition can be stated in a truth table, where *T* denotes true, and *F* denotes false, and each row states a separate possibility:

A	B	A or B, or both
T	T	T
T	F	T
F	T	T
F	F	F

The connectives of the propositional calculus can all be defined by truth tables. Strictly speaking, it is a mistake to assign truth values to sentences in natural language: The same sentence can be used to assert many different propositions; for example, the sentence "I felt ill yesterday" asserts different propositions depending on who asserts it and when it is asserted. Hence, it is

propositions, not sentences, that have truth values (Strawson, 1950).

Deductions can be made using truth tables rather than formal rules. Each premise is used to eliminate otherwise possible combinations of the atomic propositions that occur as constituents of the premises. Thus, the deduction above concerns four possibilities of two atomic propositions:

Lisa is in Cambridge	Ben is in Dublin
T	T
T	F
F	T
F	F

The first premise, Lisa is in Cambridge or Ben is in Dublin, or both, eliminates the fourth possibility in the table, which is not compatible with the truth of this premise. The second premise, Lisa is not in Cambridge, eliminates the first 2 possibilities. When you have eliminated the impossible, then whatever remains must be the case. What remains is, of course, the third possibility, in which it is true that Ben is in Dublin. This conclusion therefore follows validly from the premises, and the deduction is made solely by using the meanings of the premises to eliminate possibilities. This method needs procedures for constructing truth tables and for eliminating possibilities from them, but it does not need any formal rule of inference, such as the rule for disjunction that we described earlier.

The fact that semantic procedures do not depend on formal rules of inference can be hard to grasp. Skeptics often ask, "Where do the truth tables come from—surely one needs to know the formal rules to construct the truth tables?" The answer is that truth tables are merely a systematic way of spelling out a knowledge of the meanings of connectives. To know the meaning of an inclusive disjunction, A or B, is to know that the assertion is true if at least one of the propositions is true and that it is false only if both of the propositions are false. The skeptic, however, persists: "Surely this knowledge must derive from formal rules of inference?" In fact, the formal rules for propositional connectives are consistent with more than one possible semantics (e.g., see Kneale & Kneale, 1962, p. 678). Hence, although it is sometimes suggested that the meaning of a term derives from, implicitly reflects, or is nothing more than the rules of inference for it, this idea is unworkable (see Osherson, 1974–1976, Vol. 3, p. 253; Johnson-Laird, 1983, p. 41; Prior, 1960). On the contrary, the rules of inference must reflect the meanings of the connectives. The meaning of an assertion relates it to the world, and the meaning of a connective makes a contribution to these truth conditions. A rule of inference enables a reasoner to pass from a set of premises to a conclusion in a purely formal way, but this step is constrained by the truth conditions of the assertions.

A major part of modern logic concerns the relations between proof-theoretic methods that rely on formal rules and model-theoretic methods that rely on the meanings of expressions. Logicians have proved that any propositional inference that is valid according to the truth-table method can be derived using the formal rules of the propositional calculus. The calculus is therefore said to be *complete.* Logicians have also proved that any propositional inference that can be derived using the calculus can also be validated using truth tables. The calculus is therefore said to be *sound* (e.g., see Jeffrey, 1981). What must be emphasized, particularly in the context of psychological theories, is that the two approaches are distinct: One is syntactic and based on formal rules of inference, and the other is semantic and based on the meanings of connectives. To deny this point is to deny the significance of these proofs of completeness and soundness.

Rule Theories of Propositional Reasoning

Natural deduction has been advocated as the most plausible account of mental logic by many theorists (e.g., Braine, 1978; Braine et al., 1984; Johnson-Laird, 1975; Macnamara, 1986; Osherson, 1974–1976, 1975; Pollock, 1989; Rips, 1983, 1988; Sperber & Wilson, 1986), and at least one simulation program uses it to construct both forward and backward chains of inference (Rips, 1983). All of these theories posit an initial process of recovering the logical form of the premises. Indeed, what they have in common outweighs their differences, but here we outline three of them to enable readers to make up their own minds.

Johnson-Laird (1975) proposed a theory of propositional reasoning partly based on natural deduction. It distinguishes between primary and auxiliary rules of inference. The primary rules include the rule for disjunction presented earlier and the rule for modus ponens:

If A, then B.

A

Therefore, B.

The following is the rule introducing disjunctive conclusions:

A

Therefore, A or B, or both.

This leads to deductions that throw semantic information away; that is, the conclusion rules out fewer states of affairs than does the premise. Valid inferences that reduce information in this way, as we noted above, are not spontaneously drawn by logically untutored reasoners and strike them as odd or absurd (e.g., see Matalon, 1962). Yet, without this rule, it would be difficult to make the following inference:

If it is frosty or it is foggy, then the game will not be played.

It is frosty.

Therefore, the game will not be played.

Johnson-Laird therefore proposed that the rule (and others like it) is an auxiliary one that can be used only to prepare the way for a primary rule, such as modus ponens.

Braine and his colleagues described a series of formal theories based on natural deduction (see Braine, 1978; Braine & Rumain, 1983). Their rules differ in format from Johnson-Laird's (1975) in two ways. First, *and* and *or* can connect any number of propositions, and so, for example, the rule introducing the conjunction of premises has the following form in their theory:

$P_1, P_2, \cdots P_n$.

Therefore, P_1 and P_2 and $\cdots P_n$.

Second, Braine and his colleagues avoided the need for some auxiliary rules, such as the disjunctive rule above, by building their effects directly into the main rules. He included, for example, the rule for modus ponens:

If $P_1, P_2, \cdots$ or P_n, then Q.

P_1.

Therefore, Q.

Again, this allows for any number of propositions in the disjunctive antecedent. Sperber and Wilson (1986) also adopted this idea.

Rips (1983) proposed a theory of propositional reasoning, which he simulated in a program called ANDS (A Natural Deduction System). The rules are used by the program in the form of procedures. The program evaluates given conclusions and builds both forward- and backward chains of deduction, and therefore it maintains a set of goals separate from the assertions that it has derived. Pollock (1989) proposed a similar system in which propositional rules are systematically divided into those that can be used in forward chains and those that can be used in backward chains. In ANDS, certain rules are treated as auxiliaries that can be used only when they are triggered by a goal, as in this example:

A, B.

Therefore, A and B.

This rule could otherwise be used ad infinitum at any point in the proof to generate an unlimited series of valid conclusions. If the program can find no rule to apply during a proof, then it declares that the argument is invalid. Rips assumes that rules of inference are available to human reasoners on a probabilistic basis. We consider the evidence for rule theories later, but first we describe the model theory of propositional reasoning.

The Model Theory of Propositional Reasoning

Deductive reasoning according to the model theory depends on three main processes. First, the starting point of the deduction—verbal premises, or perceptual observations—is used to construct a set of mental models, typically a single model. In the case of verbal premises, this model is constructed on the basis of their meaning and any relevant general knowledge. In the case of observations, the end product of perception is a model of the world (Marr, 1982). Second, if no conclusion is available or provided by the experimenter, then an attempt is made to formulate one from the model: The conclusion should express information that is not directly asserted by the premises. If reasoners base their conclusions on models, then, as we point out, they are bound to maintain the semantic information in the premises. If there is no conclusion expressing something that is not explicit in the premises, then the response "nothing follows" is made. Third, if a conclusion is forthcoming, then its validity can be checked by ensuring that no model of the premises renders it false. If there is no such model, then the conclusion is valid, that is, it must be true given that the premises are true. If there is such a model, then it is necessary to return to the second stage to determine whether there is any conclusion that holds over all the models so far constructed. If it is uncertain whether such an alternative model exists, then the conclusion can be drawn on a tentative or probabilistic basis (cf. Kahneman & Tversky, 1982).

So much for the general theory. How in particular can it be applied to propositional reasoning? In fact, all that is needed is an account of how the meanings of connectives are used in the construction of models. Once this account is available, the rest of the deductive machinery is already in place to deal with the domain. The same deductive machinery can be used for all domains: What changes are the relevant terms (quantifiers, connectives, etc.) and the concomitant semantics underlying the construction of models? The question of meaning can best be approached in the following way. Because we are concerned here only with connectives for which formal rules of inference are appropriate, we assume that the connectives are truth-functional, that is, their meanings can be specified by truth tables. The mental representation of their meanings, however, is unlikely to take the form of truth tables because they are too bulky to be mentally manipulated (see later). The problem is therefore to reconcile the semantics of truth tables with the constraints of mental processing and to do so in a way that explains the phenomena of propositional reasoning. Truth tables need to be replaced with psychologically plausible models.

In essence, the model theory assumes that human reasoners represent as little information as possible explicitly and that models can contain abstract symbols that do not directly correspond to anything in the physical world. (For a defense of these symbols, see Johnson-Laird, 1983, ch. 15; Polk & Newell, 1988; and Johnson-Laird & Byrne, 1991). The more information that has to be represented explicitly, the greater the load on the processing capacity of working memory, and so the initial models of a proposition represent as much information as possible implicitly. Implicit information is not immediately available for processing, but it becomes available when it is made explicit. This process adds to the number of explicit models to be taken into account in reasoning from the premises.

We illustrate these ideas by considering the representation of various sorts of proposition. Consider the model of a conjunction, such as the following:

There is a circle and there is a triangle.

This calls for both conjuncts to be incorporated within one model of the state of affairs:

○ △

Consider next the further premise:

There is not a circle.

It cannot be added to this model without contradiction. The model may take the form of a visual image, but the crux of the theory concerns not subjective experience but the structure and number of models that are required to make inferences. The initial representation of a disjunction, such as, *There is a circle or there is a triangle,* calls for two models:

○

△

In these models we adopt the notational convention that each line in a diagram denotes a separate mental model of a different possible situation. The first line above denotes a model of the situation in which there is a circle, and the second line denotes a model of the situation in which there is a triangle. The further assertion of the categorical premise,

There is not a circle,

can be integrated with the disjunctive models in only one way. The model representing the circle must be eliminated because it is inconsistent with the premise, and the information that there is not a circle must be incorporated within the one remaining model:

¬ ○ △.

In this model ¬ denotes an abstract mental symbol representing negation. A procedure that scans models for conclusions not directly stated in premises will yield the following conclusion:

There is a triangle.

There is no alternative model of the premises that refutes this conclusion, and so it is valid. The deduction can therefore be drawn without using the formal rule of disjunctive inference and without representing the disjunction as inclusive or exclusive. The conclusion emerges from a semantic procedure that constructs and evaluates models based on the meaning of the premises. This procedure depends on rules, but, to reiterate the important point, these rules are not formal rules of inference such as modus ponens.

The initial representation of a disjunction by the models above is consistent with either an inclusive interpretation (circle or triangle, or both) or an exclusive interpretation (circle or triangle, but not both), and the models can be fleshed out explicitly to represent either sort of disjunction. The distinction depends on making explicit that all instances of a particular contingency, such as those in which there is a circle, have been exhaustively represented in the set of models. In other words, reasoners may know that there could be other instances of a circle or that they have represented all of them exhaustively. A contingency that has been exhaustively represented cannot be added to the set of models. (Strictly speaking, the notion is relative: One contingency is exhaustively represented in relation to another, but we ignore this aspect of the notion for the time being and treat the contrast as a binary one.) We use square brackets as our notation for an exhaustive representation. Thus, the exclusive disjunction,

Either there is a circle or else there is a triangle, but not both,

has the following models:

[○]

[△]

These models represent explicitly that all states containing circles and all states containing triangles have been exhausted. For example, a triangle cannot be added to the first model. Consider an inclusive disjunction, such as the following:

There is a circle or there is a triangle, or both.

This calls for three distinct possibilities:

[○] [△]

[○]

[△]

including the joint contingency of the circle and the triangle.

The same principle applies to conditionals. The initial models of a conditional, such as,

If there is a circle, then there is a triangle, are as follows:

[○] △

. . . ,

where the first model represents the joint occurrence of circle and triangle. *If* means that there may be an alternative to the situation in which the antecedent is true, and the second model depicted by the three dots allows for this possibility. This model initially has no explicit content but allows for its subsequent specification. The possibility of this alternative situation rules out any simple conjunctive description of the models. (The conjunctive state of affairs satisfies a conditional description, but it does not capture its meaning, which allows for at least one alternative possibility.) Because the triangle is not exhaustively represented, the initial models can be fleshed out in two distinct ways. One way corresponds to a conditional interpretation, and the other way corresponds to a biconditional interpretation ("If and only if there is a circle, then there is a triangle"). The conditional interpretation allows that there can be a triangle without a circle:

[○] [△]

[△]

. . .

The biconditional interpretation does not allow a triangle without a circle:

[○] [△]

. . .

Given the initial models of the conditional, the categorical premise,

There is a circle,

picks out the situation represented in the first model and eliminates the second model. The remaining model,

[○] [△]

yields the modus ponens conclusion:

There is a triangle.

Similarly, the inference from the conditional and the categorical premise,

There is a triangle,

to the conclusion,

There is a circle,

is almost equally as easy, apart from the difficulty of arguing in the opposite direction to the one in which the information from the conditional entered working memory (see Johnson-Laird & Bara, 1984; and Evans & Beck, 1981, who also argued that inferences from antecedent to consequent are easier than those in an opposite direction). This inference with a conditional is fallacious, but it is frequently made (e.g., see Evans, 1982).

The valid deduction *modus tollens* has the following sort of premises:

If there is a circle, then there is a triangle.

There is not a triangle.

If there is a circle, then there is a triangle.

There is not a triangle.

The initial models of the conditional are, as before:

[○] △

. . .

The categorical premise (there is not a triangle) eliminates the first model and adds explicit information to the second model:

[¬△]

There is no obvious conclusion to be drawn, because the model contains only the information in the categorical premise. Hence, the procedure for formulating conclusions asserts, "nothing follows." This response, as we show, is often made by subjects. In order for the modus tollens conclusion to be drawn, it is necessary to flesh out the initial models completely so that they represent the cases that do not contain circles or triangles. In the case of a conditional interpretation, the models are,

○ △

¬○ △

¬○ ¬△

Because the models are complete, we have for simplicity omitted the square brackets representing exhaustiveness. In the case of a biconditional interpretation, the models are,

○ △

¬○ ¬△

Whichever interpretation is made, the categorical premise calls for the elimination of the models containing the triangles, and so this process leaves behind only one model,

¬○ ¬△,

which yields the modus tollens conclusion:

There is not a circle.

The same conclusion is derived from the models of the biconditional. In either case, the inference is harder to draw than modus ponens, because it depends on a greater number of explicit models, which have to be fleshed out in the course of the deduction.

Negation calls for forming the complement of the relevant set of models. Thus, as we have shown, the representation of the negation,

There is not a triangle,

calls for constructing the complement of the model of the triangle:

¬△

Similarly, the negation of a conjunction,

There is not both a circle and a triangle, calls for forming the complement of the model of the conjunction:

○ △

This process consists of enumerating explicitly all the other possibilities for the two components:

○ ¬△

¬○ △

¬○ ¬△

The negation of a disjunction,

There is not either a circle or a triangle,

or, in its more familiar guise,

There is neither a circle nor a triangle,

yields the following model:

¬○ ¬△

Logically untutored reasoners spontaneously draw conclusions that maintain the semantic information of premises. Logicians define semantic information in terms of the states of affairs that a proposition rules out as false, that is, the proportion of *false* entries in the proposition's truth table[4](see Johnson-Laird, 1983, p. 36). Given two assertions, *p* and *q*, the categorical assertion of p rules out half the entries in the truth table, whereas the assertion of the conjunction, *p and q*, rules out three-quarters of the entries. The conjunction conveys more semantic information than *p*, and so a valid deduction of the form

p and *q*

$\therefore p$

throws away semantic information. The model theory assumes that reasoners formulate conclusions by framing parsimonious descriptions of the models they have constructed from the premises. They do not need, however, to include categorical premises as part of their conclusions, because there is no need to repeat the obvious (Grice, 1975). For example, consider the interpretation of these premises:

If there is a circle, then there is a triangle.

There is a circle.

The premises yield the following model:

○ △

There is no need to describe the circle, because it corresponds to the categorical premises, and so the following conclusion will be framed:

There is a triangle.

In general, the parsimonious description of a set of models is bound to maintain the semantic information in the premises.

The initial models of a premise can be highly implicit; for example,

[○] △

. . .

but they can become completely explicit about instances and their complements:

○ △
¬○ △
¬○ ¬△

Table 1 presents the initial models and the final wholly explicit models for each of the main connectives. Most inferences in daily life are probably made using initial models. They represent only those contingencies that directly correspond to the atomic propositions in the premises. In what circumstances do reasoners flesh out initial models explicitly? The principal barrier is the processing capacity of working memory, and so the following conditions must be met: (a) The premises must be simple, for example, a conditional interrelating two or three atomic propositions; and (b) nothing follows from the initial models. One factor that can assist the process of fleshing out is existing knowledge about the contingencies to be added to the models. Consider the following premises, for example:

If it was foggy, then the match was cancelled.

The match was not cancelled.

They are likely to yield the modus tollens conclusion:

It was not foggy.

People know the relations between fog and sports, and they can use this knowledge to flesh out their models of the conditional. No such existing knowledge assists the process in the case of a neutral conditional, such as,

If there is an A, then there is a 2.

Some evidence exists to support such effects of knowledge (e.g., Byrne, 1989; Cheng & Holyoak, 1985; Cummins, Lubart, Alksnis, & Rist, 1989; Griggs, 1983; Thompson, 1989). In any event, people try to use models that make explicit as little information as possible.

The relation between the models in Table 1 and truth tables should now be evident. Consider, for example, the truth table presented earlier for inclusive disjunction: A or B, or both. The explicit models of a disjunction correspond to those rows in the truth table that are true:

A B —the first line in the truth table.

A ¬B —the second line in the truth table.

¬A B —the third line in the truth table.

There is, however, a crucial distinction: Models are less bulky than truth tables. In particular, the number of entries for a truth table based on n atomic propositions is 2^n, but the number of initial models does not increase exponentially. A conjunction of n atomic propositions requires only one model, and a disjunction of n atomic propositions requires only n initial models. Osherson (1974–1976) reported only low correlations between the size of truth tables and his subjects' difficulty with arguments, but, as should now be obvious, this finding in no way impugns the model theory.

Table 1
Initial Models and the Final Wholly Explicit Models for the Main Propositional Connectives of Assertions p and q

	Model	
Connective	Initial	Explicit
p and q	p q	p q
p or q		Inclusive: p q / p $\neg q$ / $\neg p$ q
	p	Exclusive: p $\neg q$ / $\neg p$ q
	q	
If p, then q		Conditional: p q / $\neg p$ q / $\neg p$ $\neg q$
	[p] q	Biconditional: p q / $\neg p$ $\neg q$
	. . .	
p only if q	[p] q	p q / $\neg p$ q / $\neg p$ $\neg q$
	$\neg p$ [$\neg q$]	
	. . .	

Note. Each row represents an alternative model. ¬ is a symbol for negation.

A Psychological and an Artificial Intelligence (AI) Algorithm for Propositional Reasoning With Models

Most propositional deductions in daily life are simple and can be carried out without using fully explicit models. Experiments have shown, however, that subjects can make deductions from premises containing several connectives. In this section, we describe an algorithm for making such deductions and an AI extension of it that reasons with wholly explicit models. The computer programs that implement the algorithms take as input the following sort of problem:

If there is a circle or an asterisk, then there is a triangle.

There is a circle.

∴ There is a triangle.

They then determine whether the conclusion follows from the premises. For convenience, the programs will also take as input expressions using *p*, *q*, *r*, and so forth as variables denoting atomic propositions.

The process of constructing models of the premises is, in theory, informed by any relevant general knowledge, but we have not implemented this assumption. The programs construct models relying solely on a parser, a context-free grammar, and a lexicon. The lexicon specifies the syntactic category and meaning of each word in its vocabulary. The meaning of the connectives *and, or,* and *if* are functions that combine models in ways that we explain later.

The grammar contains the rules for analyzing the syntax of input premises, and each rule has an associated semantic rule so that the interpretation of a premise can be built up on a stack as the premise is parsed. The grammar contains a rule, for instance, that specifies that a single variable, such as *p*, is a well-formed sentence. This rule includes a semantic function that is evaluated whenever the rule is used by the parser, and its evaluation examines the contents of the stack and returns a model corresponding to the value of the variable. The parser accordingly builds up the models for a premise in a compositional way; that is, each time it uses a rule in the grammar in parsing a sentence, it also evaluates the corresponding semantic function. Thus, when it parses the italicized constituent,

If *there is a triangle,*

it builds a model of the clause:

△

When it parses the next clause,

then *there is a circle,*

it builds another model:

○

When it finally uses the rule that identifies the constituents,

if sentence, *then* sentence,

as themselves making up a sentence, it evaluates the rule's associated semantics. This evaluation leads to the application of the meaning of *if* to the interpretations of the two constituent sentences.

In the psychological algorithm, the meanings of the connectives, which are specified in the lexicon, generate only the initial models shown in Table 1. There is no need to represent which instances are exhaustively represented, and so the meaning of *if* yields one model formed from a conjunction of the antecedent and consequent models and then adds an implicit model:

△ ○

. . .

The meaning of *or* yields two alternative sets of models corresponding to its constituents. The meaning of *not* calls for forming the complement of a set of models. If the set of models to be negated includes an implicit model, then the set is fleshed out explicitly prior to negating it. The interpretative system has to allow for building up the representation of premises containing any number of connectives, and so the lexical meanings are, in essence, instructions for forming new sets of models out of old. For example, the interpretation of the premise,

If *c* or *h*, then *p*,

leads first to the interpretation of the disjunction,

c

h

which is then treated as the antecedent set of models in the interpretation of the conditional:

c *p*

h *p*

. . .

The meaning of *and* plays a central role in the program. It is elicited by the connective, and it is also used to combine separate premises. Two premises, such as

If *c* or *h*, then *p*.

c

are equivalent to the conjunction,

If *c* or *h*, then *p*, and *c*.

Hence, the program combines the interpretations of premises using conjunction. The semantics of *A and B,* where A and B both denote sets of models, calls for conjoining each model in A with each model in B according to the following principles:

1. If the model in A is implicit (i.e., denoted by three dots), and the model in B is implicit, then the result is an implicit model.
2. If the model in A is implicit but the model in B is explicit, or vice versa, then no new model is formed from them. The explicit model will occur in other combinations, and it does not need to be repeated (see the example below).
3. If the models are inconsistent, that is, one contains an atom and the other contains its negation, then no new model is formed from them.
4. Otherwise, the two models are joined together eliminating any redundancies.

Hence, the process of combining the two sets of models (corresponding to premises in the previous example):

c *p*

h *p* and *c*

. . .

The process proceeds as follows:

c *p* x *c* yields *c* *p*,

h *p* x *c* yields *c* *h* *p*,

. . . x *c* yields nil.

The final result is therefore

c *p*

c *h* *p*.

The high-level function controlling the building of models loops through the list of premises conjoining the models for the

current premise with the models representing all the previous premises. The problems in the Appendix illustrate the outcome of the program for a variety of propositional deductions.

One apparent anomaly arises from dropping the use of square brackets. Strictly speaking, exhaustive representation is a relative notion. In the models for "If there is a triangle, then there is a circle," the representation of the triangle is exhausted in relation to the circles, but exhaustion can be treated in this case as an absolute notion because there are no other shapes. Consider, however, the representation of the two conditionals:

If there is a triangle, then there is a circle.

If there is a cross, then there is a circle.

A completely explicit set of models, as constructed by the AI algorithm, is as follows:

$$\begin{array}{ccc} \triangle & + & \bigcirc \\ \triangle & \neg + & \bigcirc \\ \neg\triangle & + & \bigcirc \\ \neg\triangle & \neg + & \bigcirc \\ \neg\triangle & \neg + & \neg\bigcirc \end{array}$$

According to the model theory, however, human reasoners do not represent this set but rather they construct the following initial models:

$$[\triangle] \quad [+] \quad \bigcirc$$

. . .

What the exhaustion symbols (brackets) mean here is (a) the triangles are exhaustively represented in relation to the circles, (b) the crosses are exhaustively represented in relation to the circles, but (c) the triangles and crosses are not exhaustively represented in relation to one another. Hence, if a triangle occurs in a model subsequently constructed by fleshing out the implicit model, then it must be accompanied by a circle, and likewise if a cross occurs in a model subsequently constructed by fleshing out the implicit model, then it must be accompanied by a circle. The first step in fleshing out the models above is, accordingly,

$$\begin{array}{ccc} [\triangle] & [+] & \bigcirc \\ [\triangle] & & \bigcirc \\ & [+] & \bigcirc \end{array}$$

. . .

The process of fleshing out is necessary only when no conclusion follows from the initial models. It is a remarkable fact, however, that the vast majority of deductions in daily life do not require any fleshing out; for example, it is not required for any of the 61 direct reasoning problems investigated by Braine et al. (1984). Our psychological algorithm does not include a fleshing out procedure, and so it does not use exhaustion: The square brackets would be an idle wheel. However, the theory, as opposed to this computer implementation of part of it, always postulates the use of exhaustion. When the program constructs models for the pair of conditionals above, it yields the following:

$$\triangle \quad + \quad \bigcirc$$

. . .

Readers should bear in mind that this representation is a deliberate simplification and that the proper initial models should have the form,

$$[\triangle] \quad [+] \quad \bigcirc$$

. . .

The AI algorithm that we have devised is sufficiently powerful to make all and only the valid deductions in the propositional calculus (within the limitations of the computer's speed and memory). If no conclusion is presented, then the program formulates a parsimonious conclusion for itself. Because the conclusion is based on models, it automatically maintains the semantic information in the premises. The program treats each premise as though it were a potential conclusion by evaluating it in relation to the previous premises. If it follows from the previous premises, or vice versa, then the program outputs this information. The program constructs only fully explicit models. Hence, the meaning of *and* combines each model in one set with each model in the other set but eliminates inconsistent combinations and redundant atoms. The meanings of the other connectives are then defined in terms of negation and conjunction.

The advantages of the model-based procedures are twofold. First, they obviate the need to search for a formal derivation of a conclusion. Such searches call for the simulation of a nondeterministic automaton and can therefore be costly—far too time-consuming even for simple deductions that human reasoners can make in a few seconds. Second, the procedures draw conclusions from premises (as opposed to mere evaluation of given conclusions). A mark of human intelligence, as we noted, is the ability to draw conclusions that are parsimonious. The task has been avoided by most previous computer models of propositional reasoning in both psychology and artificial intelligence (e.g., see Bledsoe, 1977; Reiter, 1973; Rips, 1983). The AI program, however, contains a procedure guaranteed to construct a maximally parsimonious conclusion, that is, a description of the models that uses each atomic proposition as few times as possible (see Johnson-Laird, 1990). A solution to this problem has been proposed for descriptions that use only conjunction, disjunction, and negation (e.g., see Brayton, Hachtel, McMullen, & Sangiovanni-Vincentelli, 1984; McCluskey, 1956; Quine, 1955). Our algorithm uses all the propositional connectives, and so it can produce still more parsimonious descriptions.

Large-Scale Empirical Studies of Propositional Reasoning

A new theory should explain existing phenomena and, ideally, account for hitherto unexplained aspects of them. In this section, we show how the model theory throws light on some representative studies of propositional reasoning. In the next section, we show how the theory accounts for the principal

phenomena of propositional reasoning. In the section after that, we report the results of some experiments designed to test some novel predictions of the model theory. The theory makes two principal predictions: (a) The greater the number of explicit models to be constructed in a deduction, the harder the task should be; that is, the task should take longer, be more likely to lead to errors, and be rated as more difficult; and (b) when subjects draw conclusions for themselves, their erroneous conclusions should correspond to a proper subset of the possible models of the premises. This second prediction is of concern in the subsequent sections.

The pioneering studies of large sets of propositional deductions were carried out by Osherson (1974–1976), who examined children's and adolescents' ability to evaluate various sorts of deduction. He proposed a formal rule theory, though unlike most formalists he left open the possibility of a semantic theory. He carried out too many experiments for us to review in detail, but fortunately we can finesse such a description because we can consider instead his three main theoretical assumptions about the weighting of formal rules. His first assumption is that conjunctions are easier than disjunctions. Although many studies of concepts bear out this assumption (e.g., Bourne & O'Banion, 1971; Bruner, Goodnow, & Austin, 1956; Haygood & Bourne, 1965; Neisser & Weene, 1962), we know of no simple explanation of the phenomenon in deductive reasoning. It follows directly from the model theory, however, because a conjunction calls for only one explicit model, whereas a disjunction calls for at least two explicit models.

Osherson's (1974–1976) second assumption is that modus ponens is easier than modus tollens. Subsequent rule theories, as we show here, have offered an explanation of this difference. The model theory explains it in terms of the need to flesh out more models explicitly in the case of modus tollens.

Osherson's (1974–1976) third assumption is that negative premises cause difficulty. All theories acknowledge that negations call for additional processing during comprehension (e.g., see Clark & Clark, 1977; Wason & Johnson-Laird, 1972). However, it is important to distinguish the effects of negation on comprehension from those on reasoning. An assertion that contains, in effect, two negatives, such as,

It is false that there is not a *K* on the blackboard,

is harder to understand than the simple affirmative,

There is a *K* on the blackboard.

A double negation takes time to understand, but once it is properly understood it should have no effects on reasoning. Both of the following negative assertions similarly take time to understand:

It is not the case that there is an A or there is a B.

It is not the case that there is both an A and a B.

The first assertion, however, is equivalent to,

There is neither an A nor a B,

which corresponds to just a single model,

$$\neg A \quad \neg B$$

Hence, once you have understood the assertion, reasoning with it should be straightforward. The second assertion calls for three distinct models:

$$\begin{array}{rr} \neg A & B \\ A & \neg B \\ \neg A & \neg B \end{array}$$

Reasoning with this assertion should be difficult. The model theory therefore introduces an additional effect of negation: the number of models that negation yields.

A second major study of propositional reasoning was conducted by Rips (1983). He reported an experiment in which subjects evaluated a set of 32 deductions. The subjects' task was to assess the validity of the given conclusion for each problem. The overall performance was at chance, but Rips was able to fit his theory to the data by assessing the availability of each of his postulated rules. The correlation between the predicted and observed proportions of correct responses was high, but King (personal communication, 1989) has drawn our attention to a worrying feature of the results. Although the observed solution rates for the individual problems are fairly evenly scattered between 16% and 92%, the predicted solution rates cluster in two regions: one around 33% and the other around 75%. Only three predictions fall between 40% and 70% correct, whereas nearly a third of the observed percentages lie in this range. King commented, "In other words, the model is not correctly specified because there is extreme patterning in the residuals, which make Rips's claim to having a correct theory very suspect." There is also some doubt about whether Rips's subjects were really reasoning—always a potential problem when subjects have to evaluate given conclusions. In commenting on the subjects' failure to evaluate certain inferences correctly, Braine et al. (1984) wrote, "So high a failure rate on transparent problems suggests that the experiment often failed to engage the reasoning procedures of subjects" (p. 360).

An interesting feature of Rips's (1983) design, on which he made no comment, is that half the deductions maintained the semantic information of the premises, whereas the other half had conclusions with less semantic information than their premises, such as the following:

not-*p* and *q*.

$\therefore$ *q* and not both *p* and *r*.

King noted that some pairs of problems that Rips predicted to be equally difficult yielded large differences in their actual solution rates. For example, the following two problems were predicted to be solved on 41.2% and 40.5% of occasions, respectively:

If *p* or *q*, then not-*s*.

s

$\therefore$ not-*p* and *s*,

and

p

If *p* or *q*, then not-*r*.

$\therefore$ *p* and not both *r* and *s*.

In fact, the solution rate was 55.6% for the first problem, which maintains semantic information, but only 33.3% for the second

problem, which reduces semantic information. To throw information away is to violate one of the fundamental principles of human deductive competence, and so we can predict that performance with these problems should be poorer. Overall, the subjects correctly evaluated the 16 inferences that maintained information on 66.3% of occasions, but they correctly evaluated the 16 inferences that threw semantic information away on only 34.8% of occasions. This difference was highly significant (Kendall's $S = 202.4$ corrected for ties, $z = 3.82$; $p < .0001$, one-tailed). Only one of the problems that threw away semantic information was evaluated at better than chance level, whereas only two of the problems that maintained semantic information were evaluated at worse than chance level.

A third study of a large set of propositional deductions was carried out by Braine et al. (1984), and here we analyze their main experiment in detail. The subjects evaluated deductions about letters on an imaginary blackboard; for example,

If there is either a *C* or an *H*, then there is a *P*.

There is a *C*.

Therefore, there is a *P*.

Their task was to rate the difficulty of the deduction on a 9-point scale, where 1 signified the easiest of problems and 9 signified the hardest of problems. Some of the problems had conclusions that followed from the premises, and some of the problems had conclusions that were inconsistent with the premises. Braine and his colleagues examined three potential indices of difficulty: the length of the problem, the number of steps in a deduction according to their theory, and the difficulty weights of these steps as estimated from the data. The most striking finding was that the rated difficulty of the problems was predicted by a regression equation based on two parameters: the length of the problem and the number of steps in its derivation. The first parameter is a multiplying constant for the number of words in a problem, and it was estimated from a separate set of problems. The second parameter was a multiplying constant for the number of steps in the derivations according to the rule theory, and it was estimated from the main data. The correlation between the predicted and the obtained ratings was .79.

We consider the 61 direct reasoning problems for which the authors present complete data, and we show that the model theory also yields a reasonable fit. All of the problems can be correctly evaluated by the algorithm that uses initial models, and so we used the program to count up the number of explicit models required in order to interpret the premises. For example, consider the following problem:

If *e* or *k*, then *o*.

e and *v*

∴ Not *o*.

The first premise elicits the following models:

```
e    o
k    o
...          [2 explicit models]
```

The second premise elicits the following models:

```
e  v         [1 explicit model]
```

The result of combining them is,

```
e    o  v
e  k  o  v   [2 explicit models]
```

Hence, a total of 5 models has to be constructed to understand the premises. The final models are inconsistent with the model of the conclusion. Some sample problems are shown in the Appendix, together with the models generated by the computer program.

A multiple regression (using the BMDP stepwise program) based on the number of premise models accounted for 53% of the variance in the observed ratings. Nine of the problems contained a double-negative premise, and the addition of this variable to the regression analysis accounted for a further 11% of the variance, that is, a correlation of .80.

The model theory also makes certain new predictions, which are corroborated by Braine et al.'s (1984) results. The main prediction is that the more models that have to be constructed, the harder the deductive task should be. We tested this prediction in two main ways. First, the theory predicts the following relative difficulty of connectives *ceteris paribus* in terms of the number of initial models that they require:

and: 1 explicit model.

if: 1 explicit and 1 implicit model.

or: 2 explicit models.

not both: 3 explicit models.

The mean ratings of problems based on one premise containing one of these connectives, and one simple categorical premise, corroborated this prediction: for *and*, 1.79; for *if*, 1.88; for *or*, 2.66; and for *not both*, 3.18 (Kendall's $S = 20$; $z = 2.55$, $p < .006$).

Second, we tested whether the number of models required to interpret the premises predicted the difficulty of the ratings for the problems in which the conclusion consisted of a single atomic proposition. We excluded problems with complex conclusions or with double-negative premises, because we needed to be certain that the ratings reflected only the number of models required in interpreting the premises. We then assessed the predicted difficulty of the problems by running the computer program that constructs initial models. There are 11 problems with valid conclusions, and the correlation between the number of models to be constructed and the rated difficulty was .93 (Kendall's τ; $z = 3.8$, $p < .001$). There are 16 problems with inconsistent conclusions, and the correlation between the number of models to be constructed and the rated difficulty was .83 (Kendall's τ; $z = 4.9$, $p < .00005$). In short, the number of explicit models that have to be constructed has a highly significant relation to the rated difficulty of the deductions.

Finally, other factors do affect performance. Braine et al. (1984) found a relation between the length of a problem and the ratings, and they pointed out that what is at stake is not mere

verbosity. We suspected that the critical factor was the number of atomic propositions, that is, the number of items in a model. One model based, for example, on the conjunction of two propositions is likely to be harder to construct and to retain than one model based on a single atomic proposition. We tested this prediction with 28 problems, which were matched in pairs: Half had conclusions consisting of a single atomic proposition, and each of these problems was matched with a problem containing the same or similar premises but a conclusion based on more than one atomic proposition.[1] For example, consider the following problem:

Not both *g* and *i*.

g

∴ not *i*.

This was matched with the problem,

Not both *l* and *s*.

∴ If *l* then not *s*

The second problem should be harder, because the conclusion contains more than one atomic proposition. The overall mean ratings were 2.77 for problems with one-atom conclusions and 3.46 for problems with multiple-atom conclusions (Wilcoxon's $T = 3$; $n = 14$, $p < .001$).

The Phenomena of Propositional Reasoning

Researchers have established experimentally a number of robust phenomena of propositional reasoning (e.g., see Evans, 1982, for a review), and in this section we examine them in the light of the model theory. The first phenomenon concerns the interpretation of disjunctions and conditionals. When psychologists test whether a disjunction, such as,

There is a circle, or there is a triangle,

is interpreted inclusively or exclusively, they find that subjects do not respond in a uniform way. Typically, they are biased toward an inclusive interpretation, but a sizeable minority prefer the exclusive interpretation (Evans & Newstead, 1980; Roberge, 1978). The results are not consistent from one experiment to another, although a semblance of consistency occurs if content or context suggests one of the two interpretations (Newstead & Griggs, 1983). Conditionals yield a similar phenomenon. When content and context are neutral, an "if . . . then" premise is interpreted sometimes as a conditional and sometimes as a biconditional: Individuals are neither consistent with one another nor consistent from one occasion to another (see Evans, 1982; Staudenmayer, 1975; Staudenmayer & Bourne, 1978; Wason & Johnson-Laird, 1972). Where the context is binary, then a conditional is taken as implying its converse. Legrenzi (1970) demonstrated this point by using such conditionals as,

If the ball rolls to the left, then the red light comes on.

These conditionals were in a situation where the ball could roll either to the right or to the left and the light was either red or green.

The lack of uniformity in the interpretations of disjunctions and conditionals is puzzling because people are neither normally aware of the two possible interpretations nor of settling on one interpretation as opposed to the other. The puzzle is magnified when it is viewed through the spectacles of rule theories, because these theories presuppose an initial recovery of the logical form of premises, which presumably should make explicit that *or* is inclusive or exclusive and that *if* is a conditional or a biconditional.

The model theory readily accounts for the vagaries in the interpretations of disjunctions and conditionals. The initial models of a disjunction, as we showed earlier, are compatible with either an inclusive or an exclusive interpretation; the initial models of an "if. . . then" assertion are compatible with either a conditional or a biconditional interpretation. If there is no reason to decide whether *or* is inclusive or exclusive, or whether *if* denotes a conditional or a biconditional, they can be represented by their initial models, which even enable certain valid deductions to be drawn (e.g., see the first disjunctive inference in the section The Model Theory of Propositional Reasoning).

The second phenomenon is the difference between modus ponens and modus tollens. Deductions in the form of modus ponens,

If there is a circle, then there is a triangle.

There is a circle.

Therefore, there is a triangle,

are easier than those in the form of modus tollens:

If there is a circle, then there is a triangle.

There is not a triangle.

Therefore, there is not a circle.

Many intelligent individuals say that nothing follows in the modus tollens case (see Evans, 1982; Wason & Johnson-Laird, 1972, for reviews). Formalists explain the difference by assuming that the mind contains a rule for modus ponens but does not contain a rule for modus tollens. To carry out modus tollens, it is accordingly necessary to make a sequence of deductions. Given premises of the form,

If *p*, then *q*

not-*q*

reasoners can hypothesize *p*:

p [by hypothesis]

From this, they can use the first premise to derive:

q [by modus ponens]

[1] The pairs of problems, in Braine, Reiser, and Rumain's (1984) numbering, were as follows: valid conclusions, 4–3, 9–27, 1–37, 12–13, 7–31, 11–38, and 43–41; and inconsistent conclusions, 15–23, 17–32, 19–36, 6–39, 20–50, 42–56, and 44–57.

This conclusion, together with the second premise, yields a self-contradiction:

q and not-q [by conjunction]

The rule of reductio ad absurdum entitles the reasoner to derive the negation of any hypothesis that leads to a self-contradiction:

$\therefore$ not-p [by reductio]

Rule theories predict that the longer the derivation of a conclusion, the harder the inferential task will be. Hence, modus tollens should be harder than modus ponens.

The model theory explains the difference between modus ponens and modus tollens, too. Modus ponens depends on one explicit model, whereas modus tollens depends on two or three explicit models, depending on whether the "if. . . then" premise is interpreted as a conditional or biconditional. Unlike rule theories, the model theory also explains why this difference disappears when the conditional information is expressed using *only if* (see Evans, 1977; Evans & Beck, 1981; Roberge, 1978). For example, take the following assertion:

There is a circle only if there is a triangle.

It has the same truth conditions as another assertion,

If there is a circle, then there is a triangle,

because both are false only when there is a circle in the absence of a triangle. Braine (1978, p. 6) argued as follows:

> The behavior of *p only if q* can be explained if we try to derive the meaning of *only if* as a compound of the meanings of *only* and *if*. In ordinary usage, *only* is equivalent to a double negative or *no . . . other than* (e.g., *Only conservatives voted for Goldwater = No one other than conservatives voted for Goldwater*). We can use this equivalence to paraphrase *only* away from *p only if q*, for example, by the following steps: *p only if q = not p if other than q = if not q then not p.*

One trouble with this account is that it appears to predict a reversal in the difficulty of modus ponens and modus tollens rather than its disappearance.

According to the model theory, the assertion, "there is a circle only if there is a triangle," makes explicit two contingencies right from the start (Johnson-Laird & Byrne, 1989). Where there is a circle there has to be a triangle, and where there is not a triangle there cannot be a circle:

$$[\bigcirc] \quad \triangle$$
$$\neg\bigcirc \quad [\neg\triangle]$$
$$\ldots$$

These initial models allow both modus ponens and modus tollens to be made without any further fleshing out. Because two models are required, both deductions should be more difficult than modus ponens with an ordinary conditional. The data confirm this prediction (e.g., see Evans, 1982).

The third phenomenon concerns disjunction: When people reason from a disjunction and a categorical premise, the task is easier with an exclusive disjunction than with an inclusive disjunction. Newstead and Griggs (1983, p. 97) argued that exclusive disjunctions are straightforward because the deductions are symmetrical: The truth of one component implies the falsity of the other, and vice versa. It is not clear, however, why such a symmetry should make deductions easier. The model theory suggests a simple alternative explanation: Exclusive disjunctions call for a smaller number of explicit models than inclusive disjunctions.

Finally, one well-established phenomenon is that different contents can exert a qualitative effect on the nature of the inferences that subjects draw (e.g., see Byrne, 1989; Cheng, Holyoak, Nisbett, & Olivier, 1986; Griggs & Cox, 1982; Wason, 1983; Wason & Johnson-Laird, 1972). Space does not permit us to consider these effects here, other than to make one observation. Because formal rules of inference are, by definition, blind to content, the only way in which rule theories can explain these effects is in terms of the initial interpretation of the premises (e.g., see Braine & Rumain, 1983; Henle, 1962) or as a result of censorship following the deductive process. As Manktelow and Over (1987) point out, however, the effects cannot be satisfactorily explained in this way. We show elsewhere that they can be accounted for by the model theory (Johnson-Laird & Byrne, 1991).

Experimental Tests of the Model Theory

A new theory should suggest new phenomena. The present theory does indeed lead to some novel predictions, and we report the results of four experiments designed to test them. In all four of the experiments, we tested the predictions of the model theory about the conclusions that subjects spontaneously draw from verbal premises.

Experiment 1: Conditional and Disjunctive Deductions

The model theory predicts that it should be harder to reason from an exclusive disjunction, such as,

Linda is in Amsterdam or Cathy is in Majorca, but not both,

than to reason from a conditional, such as,

If Linda is in Amsterdam, then Cathy is in Majorca.

The exclusive disjunction yields two explicit models: one representing Linda in Amsterdam, and the other representing Cathy in Majorca, whereas the conditional yields, at least initially, one explicit model. Hence, in general, deductions based on exclusive disjunctions should be harder to make than those based on conditionals, because disjunctions from the outset place a greater load on the capacity of working memory: They demand an immediate representation of two explicit models, whereas the conditionals initially require only one. Some corroboratory evidence already exists in the literature. Roberge (1978), for example, obtained such an effect, but his study was limited to only one sort of deduction. Evans and Newstead (1980) similarly report that when one constituent of a conditional is negated, reasoners can still cope, but they become hopelessly lost when one constituent of a disjunction is negated. The present experiment was designed to make a systematic comparison of deductions based on exclusive disjunctions with those based on

conditionals. Consider, for example, a deduction based on the following premises:

Linda is in Amsterdam, or Cathy is in Majorca, but not both.

Linda is in Amsterdam.

The deduction has an affirmative categorical premise and implies the following conclusion:

Cathy is not in Majorca.

It initially requires two explicit models. The analogous affirmative deduction based on a conditional is, of course, a case of modus ponens, and it initially requires one explicit model. A deduction based on such premises as,

Linda is in Amsterdam, or Cathy is in Majorca, but not both.

Cathy is not in Majorca.

has a negative categorical premise and implies the following conclusion:

Linda is in Amsterdam.

It again requires initially two explicit models. The analogous negative deduction with a conditional is modus tollens, and it initially requires an explicit model and an implicit model, which is then fleshed out with one or two further explicit models. The number depends on whether the assertion is interpreted as a conditional or a biconditional.

We can predict that the negative deductions should be harder overall than the affirmative deductions, because the negative deductions require an inconsistency to be detected between the model of the categorical premise and the model for one of the disjuncts. In principle, a negative deduction with a conditional calls for two or three models to be made explicit, whereas with the disjunction it calls for only two models to be made explicit. This process of fleshing out occurs after the initial interpretations of the main premises, however, and so reasoners should already have run into trouble with the disjunctions. Should the two variables interact? It is not entirely clear. The difference between the two conditional deductions should be relatively large (one model vs. two or three models), whereas the only difference between the disjunctive deductions is that the negative inference calls for detecting an inconsistency. However, the two factors are not commensurable.

Method. The subjects acted as their own controls and carried out four deductions in each of four conditions: affirmative conditional (modus ponens), negative conditional (modus tollens), affirmative disjunction, and negative disjunction. The 16 experimental trials occurred in a different random order for each subject, and in addition there were a further six filler items that occurred in fixed positions throughout the experiment. The lexical content of the problems concerned people and well-known places (as in the examples above). We used an equal number of male and female proper names, and no place name occurred in more than one deduction. We devised two sets of such materials and assigned them at random to the problems in two different ways. Half the subjects received one set of materials, and the other half received the other set of materials. Hence, no subject encountered a particular content more than once.

The subjects were tested individually. Each premise of a problem was printed on a separate sheet of paper. The subjects read aloud the first premise, and when they were ready, they were given the second premise to read aloud. Finally, they wrote down what conclusion, if any, they thought followed from the premises. They were told that they could take as much time as they wanted for each problem. They were given a single practice trial prior to the start of the experiment proper.

We tested 14 members (10 female and 4 male) of the Medical Research Council (MRC) Applied Psychology Unit, Cambridge, subject panel, whose ages ranged from 24 to 60 years. These subjects came from a variety of occupations and were more representative of the population at large than university students. We paid them £3.60 per hr for participating in the experiment, which lasted for about 15 min. We rejected 3 of the subjects prior to the analysis of the data because they could not perform the task.

Results and discussion. The percentages of correct conclusions to the deductions were as follows: 91 for affirmative conditionals, 64 for negative conditionals, 48 for affirmative disjunctions, and 30 for negative disjunctions. Overall, as we predicted, the conditional deductions were easier than the disjunctive deductions for every subject, apart from three ties ($p = .5^8$). Likewise, the affirmative deductions were easier than the negative deductions for every subject, apart from two ties ($p = .5^9$). Although the trend suggested an interaction between these two variables, it was not quite significant (Wilcoxon's $T = 12$, $n = 7$, $p > .05$). However, the difference between the two sorts of conditional deductions was significant (Wilcoxon's $T = 5.5$, $n = 8$, $p < .05$), whereas the difference between the two sorts of disjunctive deductions was not (Wilcoxon's $T = 7.0$, $n = 8$, $p > .05$).

The experiment confirmed the critical prediction that conditional inferences would be easier than disjunctive inferences. Hence, the number of explicit models that have to be constructed ab initio does appear to affect the difficulty of a deduction, and, similarly, the need to detect an inconsistency also increased the difficulty of the task. The possible prediction of an interaction was not confirmed, though the trend was in the correct direction. A more powerful experiment might have yielded a reliable result.

Experiments 2 and 3: Conditional and Biconditional Deductions

The model theory predicts that modus ponens should be equally easy whether the main premise is a conditional or a biconditional, but that modus tollens should be easier with a biconditional than with a conditional. Modus tollens requires explicit models, and for the biconditional,

If and only if there is a circle, then there is a triangle,

they are as follows:

○ △

¬○ ¬△

The corresponding conditional, however, has these fully explicit models:

○ △

¬○ △

¬○ ¬△

In both cases, the correct response is the same. Given the categorical premise,

There is not a triangle,

it follows from both conditional and biconditional that,

∴ There is not a circle.

However, modus tollens should be easier with a biconditional, which requires two explicit models, than with a conditional, which requires three explicit models. The simpler modus ponens deduction can be made from the single initially explicit model whether the premise is a conditional or a biconditional. Experiment 2 tested these predictions.

Method. The subjects acted as their own controls and carried out two inferences in each of eight different conditions. The conditions depended on whether a deduction was in the form of modus ponens or modus tollens, whether the first premise was a conditional (*if _ then _*) or a biconditional (*if and only if _ then _*), and, for the sake of a varied set of problems, whether there were two premises, as in the example above, or three premises, such as the following:

If Mary is in Dublin, then Joe is in Limerick.

If Joe is in Limerick, then Lisa is in Princeton.

Mary is in Dublin.

What follows?

Half the subjects received all the deductions based on conditionals and then all the deductions based on biconditionals, and the other half received the two blocks in the opposite order. Within each block, the order of the deductions was randomized for each subject.

The modus tollens deductions were based on premises containing contrary atomic propositions, for example,

If Mary is in Dublin, then Joe is in Limerick.

Joe is in Cambridge.

What follows?

We deliberately used these inconsistencies rather than negations to keep the design of the experiment simple; that is, we could compare the deductions without having to worry about the location of the negation (in the conditional or the categorical premise).

The lexical contents of the problems referred to the locations of people in well-known cities. As in the previous experiment, we used an equal number of male and female proper names, and no city occurred in more than one deduction. We randomly assigned the lexical contents to the problems in four ways, with the restriction that any items assigned to a conditional in one set were assigned to a biconditional in a different set. The resulting sets of materials were assigned to the subjects at random.

The subjects were tested individually and given two practice problems based on quantifiers. They were asked to state what, if anything, followed from each set of statements. If they considered that nothing followed, then they were to say so. The problems were printed on separate sheets of paper, and the subjects gave their responses orally. They were under no time pressure.

We tested 16 subjects (14 female and 2 male) from the subject panel of the MRC Applied Psychology Unit, whose ages ranged from 22 to 57 years. We paid them £3.60 per hr for taking part in the experiment, which lasted for about 15 min.

Results, replication, and discussion. Table 2 presents the percentages of correct conclusions drawn by the subjects for each sort of deduction. We have collapsed the data from the two different orders of presentation of the two blocks of deductions because there was no reliable difference in accuracy between them. Overall, modus ponens was reliably easier than modus tollens (Wilcoxon's $T = 1.5$, $n = 12$, $p < .005$, one-tailed); two-premise problems were marginally easier than three-premise problems (Wilcoxon's $T = 3.0$, $n = 7$, $p < .05$, one-tailed); and there was no reliable difference between conditional and biconditional problems (Wilcoxon's $T = 9.5$, $n = 8$). Although the trend was in the right direction, the overall interaction between the type of deduction and the type of conditional was not significant (Wilcoxon's $T = 6.5$, $n = 8$). However, this interaction was significant for the two-premise problems: the difference between modus ponens and modus tollens was larger for conditionals than for biconditionals (Wilcoxon's $T = 0$, $n = 6$, $p < .005$). Indeed, modus tollens with a biconditional was significantly easier than with a conditional (Wilcoxon's $T = 0$, $n = 5$, $p < .005$).

Table 2
Percentages of Correct Deductions Made in Experiments 2 and 3

	Modus Ponens		Modus Tollens	
Condition	Exp 2	Exp 3	Exp 2	Exp 3
Two-premise				
Conditional	97	96	38	56
Biconditional	97	98	59	67
Three-premise				
Conditional	88	96	38	42
Biconditional	84	100	44	65

Note. Exp = experiment.

Although the results were promising, the failure to obtain an overall interaction led us to carry out Experiment 3, which was a replication of Experiment 2, but with increased power. We tested 24 volunteer students at the University of Leuven (15 female and 9 male) whose ages were 17 to 18 years. The percentages of correct conclusions are also shown in Table 2. Once again, we have collapsed the data from the two different orders of presentation of the deductions because there was no reliable difference in accuracy between them. Overall, modus ponens was reliably easier than modus tollens for every single subject (Wilcoxon's $T = 0$, $n = 17$, $p < .005$, one-tailed); and the biconditional problems were 10% easier than the conditional problems (Wilcoxon's $T = 7.5$, $n = 10$, $p < .025$). The main finding, however, was that the overall interaction between the type of deduction and the type of conditional was significant (Wilcoxon's $T = 7.5$, $n = 10$, $p < .025$). There was no reliable difference between the two sorts of conditional for modus ponens (Wilcoxon $T = 1$, $n = 2$, *ns*), but modus tollens was reliably easier with a biconditional than with a conditional (Wilcoxon $T = 4.5$, $n = 10$, $p < .01$).

The experiments confirmed the prediction that the greater the number of explicit models required to make a deduction, the harder the deduction is to make. The use of a conditional or biconditional premise did not affect modus ponens, but the subjects were more likely to draw the correct modus tollens conclusion with a biconditional than with a conditional. The effect of the form of the argument (modus ponens vs. modus tollens) was greater than the effect of the type of the condi-

tional (conditional vs. biconditional). The difference is to be expected, however, because modus tollens also calls for the detection of an inconsistency between the two sets of models.

Experiment 4: Double Disjunctions

A major prediction of the model theory that we have not so far examined is that erroneous conclusions should correspond to some of the possible models of the premises. If the theory is correct, then as soon as the number of models that reasoners have to construct exceeds the capacity of their working memories, they are likely to be unable to reach a correct conclusion. The difficulty should be particularly exacerbated by disjunctive premises. If you wish to experience this phenomenon, then ask yourself what, if anything, follows from these premises:

June is in Wales, or Charles is in Scotland, but not both.

Charles is in Scotland, or Kate is in Ireland, but not both.

Each premise calls for two explicit models, but when the two sets are combined they yield only two models:

[W] [I]

[S]

where *W* denotes June in Wales, *I* denotes Kate in Ireland, and *S* denotes Charles in Scotland. The two models support the following conclusion:

June is in Wales, and Kate is in Ireland,

or Charles is in Scotland, but not both.

Exclusive disjunctions should be easier to cope with than inclusive disjunctions, such as these:

June is in Wales, or Charles is in Scotland, or both.

Charles is in Scotland, or Kate is in Ireland, or both.

This is because each of these premises calls for three explicit models that combine to yield five models:

[W] [S] [I]

[W] [S]

[W] [I]

[S] [I]

[S]

The models support the following conclusion:

June is in Wales, and Kate is in Ireland,

or Charles is in Scotland, or both.

Hence, a double exclusive disjunction should be reliably easier than a double inclusive disjunction.

The aim of the present experiment was, in part, to test this prediction, and we also compared negative deductions in which a constituent and its contrary occur in the two premises; for example,

June is in Wales, or Charles is in Scotland, or both.

Charles is in England, or Kate is in Ireland, or both.

Of course, if Charles is in Scotland, then he is not in England. According to the model theory, affirmative deductions should be easier than such negative deductions because the latter call for the detection of the inconsistency between the contrary constituents.

Exclusive affirmative:	2 models per premise	2 final models
Exclusive negative:	2 models per premise	3 final models + inconsistency
Inclusive affirmative:	3 models per premise	5 final models
Inclusive negative:	3 models per premise	5 final models + inconsistency

The principal goal of the experiment, however, was to examine the nature of the erroneous conclusions that the subjects inferred. Any deduction that calls for three or more models should be very difficult, and so we expected that the subjects would make many errors. The key question was, would the errors correspond to a proper subset of the possible models of the premises?

Method. The subjects acted as their own controls and carried out two deductions in four conditions based on inclusive or exclusive disjunctions and an affirmative or negative relation between the constituents common to the two premises. In addition to the experimental materials, there were 16 filler items based on simple disjunctive deductions. Half the subjects carried out all of the deductions based on exclusive disjunctions and then all of the deductions based on inclusive disjunctions, and half the subjects received the two blocks in the opposite order. The order of the problems (including the filler items) within the blocks was randomized for each subject.

The lexical materials again concerned people in cities, and we assigned them twice at random to the forms of problems, with the constraint that items assigned to an exclusive disjunction in one set were assigned to an inclusive disjunction in the other set. Half the subjects were tested with one set of materials, and half the subjects were tested with the other set of materials.

The procedure was the same as in the previous experiments: The subjects were asked to say what, if anything, followed from the premises. The subjects were tested individually. The difference between an inclusive disjunction and an exclusive disjunction was explained to them, and they were given two practice problems. The problems were printed on separate sheets of paper, and the subjects gave their responses orally.

We tested 24 subjects (16 female and 8 male) from the subject panel of the MRC Applied Psychology Unit. Their ages ranged from 18 to 59 years. They were paid £3.60 per hr for taking part in the experiment, which lasted approximately 20 min.

Results and discussion. The percentages of valid conclusions to the four sorts of deduction were as follows:

Exclusive affirmative:	21
Exclusive negative:	8
Inclusive affirmative:	6
Inclusive negative:	2

We have collapsed the results from the different orders of presentation because they had no effect on accuracy. As the model

theory predicted, exclusive disjunctions were reliably easier than inclusive disjunctions (Wilcoxon's $T = 4.5$, $n = 10$, $p < .01$), and the affirmative problems were reliably easier than the negative ones (Wilcoxon's $T = 2$, $n = 6$, $p < .04$). As we expected, the task was very difficult and there was a floor effect: Once a deduction called for three models, it became almost impossible for our subjects (see Johnson-Laird & Bara, 1984, for the same effect in syllogistic reasoning).

The subjects drew many erroneous conclusions. Figure 1 presents the percentages of erroneous conclusions according to their consistency with one or more models of the premises. Most conclusions that combine the atomic propositions in the problems with connectives are not compatible with a subset of the premise models (discussed shortly). Yet, as the figure shows, nearly all the subjects' erroneous conclusions can be accounted for by the model theory: Only a small percentage were not consistent with any subset of the models of the exclusive disjunctions, and there were none whatsoever for the inclusive disjunctions. Indeed, the figure shows a striking peak in the results for all four sorts of deduction: The most frequent category of response was a conclusion based on just a single model of the premises. For example, the double disjunction,

June is in Wales or Charles is in Scotland, but not both.

Charles is in Scotland or Kate is in Ireland, but not both.

elicited the following typical error:

June is in Wales, and Kate is in Ireland.

A small proportion of conclusions were consistent with more than five models: They consisted of logically very weak conclusions, such as "If Linda is in Amsterdam, then maybe Cathy is in Palermo and maybe Fiona is in Stockholm." They appeared to have been constructed not by examining all of the models but rather by making a prudent qualification with a word such as *maybe* of a conclusion based on a subset of models.

The only responses that are not shown in the figure were of the form "no valid conclusion," and they occurred overall on just under a third of the trials (exclusive affirmative 23%, exclusive negative 29%, inclusive affirmative 25%, and inclusive negative 44%). As we saw earlier, the model theory proposes that the response, "no valid conclusion," is made whenever the models of the premises fail to support any conclusion that is both novel and parsimonious, for example, in the case of the initial models for a modus tollens argument. Naturally, if reasoners are unable to construct a set of models or are unable to discern what holds over all of them, they will also make the same response. Hence, the response is more frequent when subjects work under time pressure: The response increased reliably in a study of syllogisms when there was a limit of 10 s in which to try to draw a conclusion (Johnson-Laird & Bara, 1984).

Could the subjects' erroneous conclusions be the result of guessing? The answer is negative because of the sheer improbability of guessing a conclusion that would correspond to a subset of the premise models. For example, there are eight distinct ways in which three individuals could be located, but only two of them are consistent with the exclusive affirmative premises. Moreover, if subjects guess a conclusion that combines two atomic propositions with a connective—and many errors had

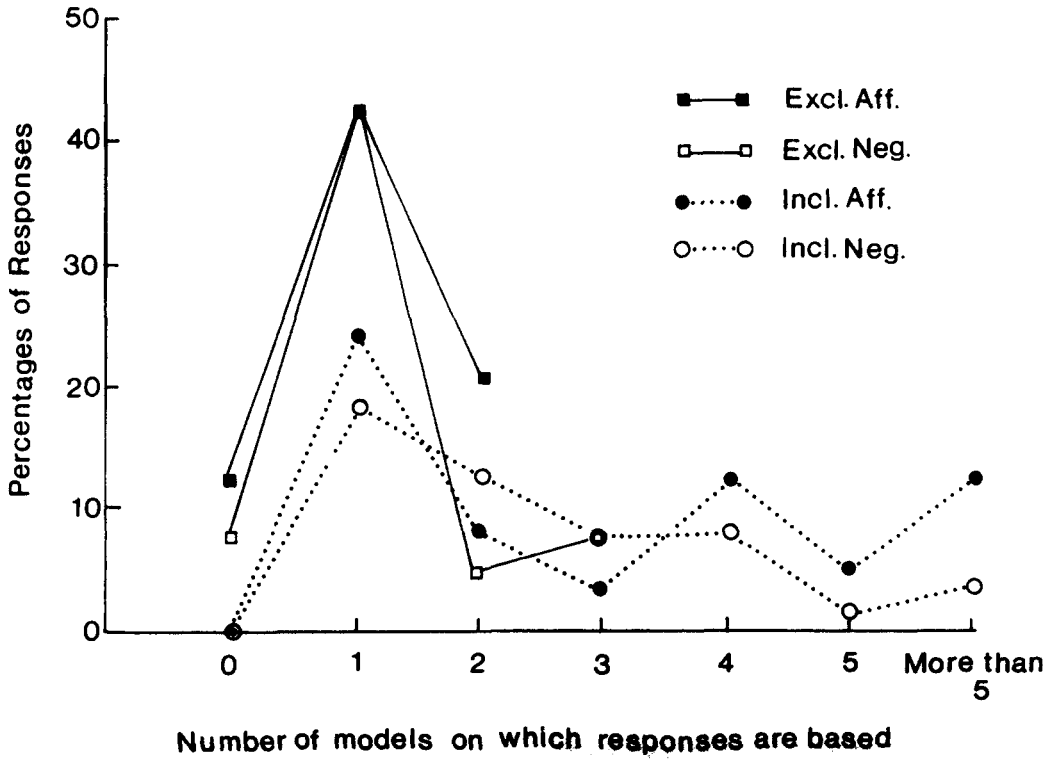

Figure 1. The percentages of responses in Experiment 4 as a function of the number of models of the premises with which they were consistent. (Excl. Aff. = exclusive affirmative; Excl. Neg. = exclusive negative; Incl. Aff. = inclusive affirmative; Incl. Neg. = inclusive negative.)

that form—then they would need to guess the propositions and the connective. If we assume that the subjects' guesses never include negations and that they restrict their connectives to truth-functional ones, then the probability of guessing a conclusion that is a subset of the premise models for the exclusive affirmatives is less than a third. Hence, two-thirds of the erroneous conclusions should not correspond to any subsets of the models; in fact less than 15% of the conclusions fell into this category (see Figure 1).

The model theory predicts a simple way in which to prevent subjects from being swamped by a double disjunction: They should be given a categorical premise in addition to the two disjunctions; for example,

Mary is in Oxford.

Mary is in Dublin, or Joe is in Limerick, or both.

Joe is in Cambridge, or Ann is in Galway, or both.

What will then happen is that the interpretation of the first 2 premises will yield only a single model:

O L

where *O* represents Mary in Oxford and *L* represents Joe in Limerick. The combination of this model with the models for the third premise yields the following:

O L G

where *G* represents Ann in Galway. Hence, the number of models that have to be kept in mind at any one time is much reduced in such a deduction. Our experiment included such conditions, which constituted the filler items, and the task was indeed very much easier in this case.

One salient aspect of double disjunctions is that the derivation of a conclusion according to formal rules is relatively straightforward and calls for no more steps than the derivation of the modus tollens deductions in Experiment 1. Given two inclusive disjunctions of the form,

p or *q*: Mary is in Dublin, or Joe is in Limerick, or both.

r or *s*: Joe is in Cambridge, or Ann is in Galway, or both.

and the premise establishing that Joe cannot be in both places,

If *q*, then not *r*: If Joe is in Limerick,
then he is not in Cambridge.

a formal derivation proceeds as follows:

Not *p*	[by hypothesis]
q	[disjunctive rule from the first premise]
Not *r*	[modus ponens from the third premise]
s	[disjunctive rule from the second premise]
If not *p*, then *s*	[Conditional proof]

Part of the difficulty of a double disjunction might be that subjects rarely reason hypothetically from any premises apart from conditionals (see also Braine et al., 1984). Yet, it is difficult to see how a rule theory could account for the fact that the erroneous conclusions tend to correspond to subsets of the premise models. Existing rule theories have yet to address this problem, presumably because they have almost always been tested by asking subjects to evaluate given conclusions. In contrast, the model theory predicts the models that can be built from the premises: Valid conclusions correspond to all the models, but when there are many models reasoners are likely either to respond that there is no valid conclusion or to draw an erroneous conclusion on the basis of a subset of the models.

General Discussion

The model theory accounts for the principal phenomena of propositional reasoning. It explains the vagaries in the interpretation of conditionals and disjunctions, the greater ease of modus ponens over modus tollens, and the disappearance of this difficulty for *only if* assertions. The model theory has also led to a number of novel phenomena. It predicts the following rank order of difficulty of connectives in simple deductions: *and, if, or,* and *not both;* this prediction was borne out by Braine et al.'s (1984) data. The model theory predicts that deductions based on conditionals should be easier than those based on exclusive disjunctions, that modus tollens with a biconditional should be easier than modus tollens with a conditional, and that deductions from exclusive disjunctions should be easier than deductions from inclusive disjunctions. Most important, the theory predicts that erroneous conclusions should correspond to subsets of possible models of the premises. Our experiments have corroborated all of these predictions.

Although it is tempting to regard the experiments as providing a crucial comparison between the model theory and theories based on formal rules, we resist the temptation. On the one hand, none of the evidence rules out a theory based on both models and rules: Only parsimony could count against a combination of both these modes of reasoning. On the other hand, a rule theory might be developed that would be consistent with our findings. Because the form of rules is not constrained, rule theories have the power of a universal Turing machine, and so they should be able to accommodate any results. Indeed, as we show here, some authors argue that the model theory is just another sort of formal rule theory.

The task of framing a rule theory that explains our results will not be easy—at least if the theory is within the confines of existing rule theories, that is, operating on the logical form of premises. The lengths of the formal derivations of simple deductions from *if* and *or* are likely to be identical—they call for only a single application of a rule. Hence, a rule theory could account for the difference in difficulty only in terms of the ease of use (or availability) of the rules of inference. This factor, however, is precisely the one that has to be assessed from data in current rule theories. Similarly, the finding that erroneous conclusions correspond to subsets of models of the premises is hard to reconcile with rule theories. Errors are supposed to arise from the failure to retrieve, or to use properly, a formal rule that is needed in a derivation. Such a spanner in the works can hardly explain the pattern of errors. Indeed, rule theories are placed in jeopardy by any fallacious conclusions that cannot be the result of misinterpretations of the premises.

What are the differences between reasoning by rule and reasoning by model? We raise the question because some authors, such as Goldman (1986, p. 292) and Rips (1990), deny the existence of any differences. The distinction between formal

rules and mental models collapses, they say, because of the abstract nature of the procedures for constructing models. Indeed, the following represents an explicit set of models for, say, the premise, "There is either a circle, or there is a triangle, or both":

○ △
○ ¬△
¬○ △

This premise is isomorphic to an expression in so-called *disjunctive normal form* (DNF), which consists of a disjunctive combination of a series of conjunctions:

There is a circle, and there is a triangle,
or there is a circle, and there is not a triangle,
or there is not a circle, and there is a triangle.

Hence, in principle, it is possible to mimic the model theory of propositional reasoning by framing rules that convert each premise into DNF (with initial implicit clauses) and then carry out the operations equivalent to those postulated by the model theory. The computer program that models our theory is precisely such an emulation, because no existing programs have any real grasp of meaning, that is, the truth conditions of assertions.

Three points need to be made about a formal theory that mimics the model theory. First, such a theory would have entirely different empirical consequences from orthodox rule theories. These theories use representations of the logical form of premises, and the process of deduction consists of the application of natural deduction rules to these logical forms in a search for a derivation that leads from premises to conclusion (see the quotation from Rips, 1983, in our introduction). The model theory does not need to search for a derivation.

Second, no existing rule theory remotely resembles the reconstruction of model theory within a framework of disjunctive normal forms: DNF is not a plausible linguistic representation of the logical form of sentences. Moreover, the model theory postulates representations that transcend truth-functional connectives; for example, it readily incorporates connectives that are not truth-functional, such as *before* and *after*. Indeed, a central distinction between the two sorts of theories is that those based on rules postulate different formal rules for different connectives, whereas the model theory postulates different meanings for different connectives. To accommodate a new connective, a rule theory must provide appropriate formal rules of inference governing its use. If the theory is also to account for how the term is understood, then it must also provide a semantic analysis of its meaning. These two accounts must be compatible with one another, but, as we argued in the introduction, the semantic analysis cannot be reduced to a set of formal rules of inference. For the model theory to accommodate a new connective, however, it needs only an account of its semantics. Its existing inferential procedures will then immediately extend to the new connective.

Third, human beings genuinely understand the meaning of assertions, and the model theory postulates that they use this ability to make deductions (cf. Johnson-Laird, 1983, p. 399). Semantics cannot be reduced to formal rules. The very idea is akin to mistaking proof theory for model theory—a mistake that renders proofs of completeness vacuous. Indeed, the original formalist program in logic was brought to an end by Gödel's celebrated incompleteness proofs. Some theorists (e.g., Penrose, 1989) argue that these proofs show that human semantic competence cannot be a computational matter. That an abstract computational device, such as a Turing machine, has no grasp of semantics seems self-evident, but robots can represent the world, and in principle they could be equipped with symbolic methods of communication that rely on an underlying semantics. What is crucial for the present argument, however, is that human beings can understand the meaning of connectives and that this process cannot consist of transforming a premise into disjunctive normal form. DNF is merely another linguistic expression, which in turn would need a semantic interpretation. The model theory assumes that human beings are conceptually equipped to envisage alternative situations—to construct alternative models—and that they learn how the semantics of connectives relates to sets of these envisaged alternatives.

In conclusion, the evidence challenges existing rule theories, but it is accounted for by the model theory. This theory is in principle simple to refute: An easy deduction that depends on many models violates its principal prediction. Yet, as we mentioned at the outset, the theory also accounts for the phenomena of relational reasoning, reasoning with single quantifiers, and reasoning with multiple quantifiers. The model theory has thus been corroborated in all of the main domains of deductive reasoning, whereas there are as yet no psychological theories based on formal rules for reasoning with quantifiers.

References

Bledsoe, W. W. (1977). Non-resolution theorem proving. *Artificial Intelligence, 9,* 1–35.

Boole, G. (1948). *The mathematical analysis of logic, being an essay towards a calculus of deductive reasoning.* Oxford, England: Basil Blackwell. (Original work published 1847)

Bourne, L. E. Jr., & O'Banion, K. (1971). Conceptual rule learning and chronological age. *Developmental Psychology, 5,* 525–534.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85,* 1–21.

Braine, M. D. S., Reiser, B. J., & Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. *The psychology of learning and motivation* (Vol. 18). San Diego, CA: Academic Press.

Braine, M. D. S., & Rumain, B. (1983). Logical reasoning. In J. H. Flavell & E. M. Markman (Eds.), *Carmichael's handbook of child psychology: Vol. 3. Cognitive Development* (4th ed.). New York: Wiley.

Brayton, R. K., Hachtel, G. D., McMullen, C. T., & Sangiovanni-Vincentelli, A. L. (1984). *Logic minimization algorithms for VLSI synthesis.* New York: Kluwer.

Bruner, J. S., Goodnow, J. J., & Austin, G. (1956). *A study of thinking.* New York: Wiley.

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition, 31,* 61–83.

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

Cheng, P. N., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17,* 391–416.

Cheng, P. N., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18,* 293–328.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics.* New York: Harcourt Brace Jovanovich.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1989). *Conditional reasoning and causation.* Unpublished master's thesis, Department of Psychology, University of Arizona, Tucson.

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

Evans, J. St. B. T. (1977). Linguistic factors in reasoning. *Quarterly Journal of Experimental Psychology, 29,* 297–306.

Evans, J. St. B. T. (1982). *The psychology of deductive reasoning.* London: Routledge, Chapman & Hall.

Evans, J. St. B. T. (1987). Reasoning. In H. Beloff & A. M. Colman (Eds.), *Psychological survey* (Vol. 6). Winchester, MA: Allen & Unwin.

Evans, J. St. B. T., & Beck, M. A. (1981). Directionality and temporal factors in conditional reasoning. *Current Psychological Research, 1,* 111–120.

Evans, J. St. B. T., & Newstead, S. E. (1980). A study of disjunctive reasoning. *Psychological Research, 41,* 373–388.

Goldman, A. I. (1986). *Epistemology and cognition.* Cambridge, MA: Harvard University Press.

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

Griggs, R. A. (1983). The role of problem content in the selection task and THOG problem. In J. St. B. T. Evans (Ed.), *Thinking and reasoning: Psychological approaches.* London: Routledge, Chapman & Hall.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic materials effect in Wason's selection task. *British Journal of Psychology, 73,* 407–420.

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

Haygood, R. C., & Bourne, L. E. Jr. (1965). Attribute- and rule-learning aspects of conceptual behavior. *Psychological Review, 72,* 175–195.

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69,* 366–378.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* London: Routledge, Chapman & Hall.

Jeffrey, R. C. (1981). *Formal logic, its scope and limits* (2nd. ed.). New York: McGraw-Hill.

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

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. (1990). *Propositional reasoning: An algorithm for deriving parsimonious conclusions.* Unpublished manuscript.

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

Johnson-Laird, P. N., & Byrne, R. M. J. (1989). *Only* reasoning. *Journal of Memory and Language, 28,* 313–330.

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

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

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge, England: Cambridge University Press.

Kneale, W., & Kneale, M. (1962). *The development of logic.* Oxford, England: Clarendon.

Legrenzi, P. (1970). Relations between language and reasoning about deductive rules. In G. B. Flores D'Arcais & W. J. M. Levelt (Eds.), *Advances in psycholinguistics* Amsterdam: North-Holland.

Levesque, H. J. (1986). Making believers out of computers. *Artificial Intelligence, 30,* 81–108.

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

Manktelow, K. I., & Over, D. E. (1987). Reasoning and rationality. *Mind and Language, 2,* 199–219.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York: Freeman.

Matalon, B. (1962). Etude genetique de l'implication [Genetic study of implications]. In E. W. Beth et al. *Implication, formalisation et logique naturelle.* Paris: Presses Universitaires de France.

McCluskey, E. J. (1956). Minimization of Boolean functions, *Bell Systems Technical Journal, 35,* 1417–1444.

Neisser, U., & Weene, P. (1962). Hierarchies in concept attainment. *Journal of Experimental Psychology, 64,* 640–645.

Newell, A. (1981). Reasoning, problem solving, and decision processes: The problem space as a fundamental category. In R. Nickerson (Ed.), *Attention and performance* (Vol. 8). Hillsdale, NJ: Erlbaum.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

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

Newstead, S. E., & Griggs, R. A. (1983). The language and thought of disjunction. In J. St. B. T. Evans, (Ed.), *Thinking and reasoning: Psychological approaches* (pp. 76–106). London: Routledge, Chapman & Hall.

Osherson, D. N. (1974–1976). *Logical abilities in children* (Vols. 1–4). Hillsdale, NJ: Erlbaum.

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

Penrose, R. (1989). *The emperor's new mind: Concerning computers, minds, and the laws of physics.* London: Oxford University Press.

Polk, T., & Newell, A. (1988). Modelling human syllogistic reasoning in SOAR. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

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

Prior, A. N. (1960). The runabout inference-ticket. *Analysis, 21,* 38–39.

Quine, W. V. (1955). A way to simplify truth functions. *American Mathematical Monthly, 59,* 521–531.

Reiter, R. (1973). A semantically guided deductive system for automatic theorem-proving. *Proceedings of the Third International Joint Conference on Artificial Intelligence* (pp. 41–46).

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

Rips, L. J. (1986). Mental muddles. In M. Brand & R. M. Harnish (Eds.), *Problems in the representation of knowledge and belief* (pp. 258–286). Tucson, AZ: University of Arizona Press.

Rips, L. J. (1988). Deduction. In R. J. Sternberg & E. E. Smith (Eds.), *The psychology of human thought.* Cambridge, England: Cambridge University Press.

Rips, L. J. (1990). Reasoning. *Annual Review of Psychology, 41,* 321–353.

Roberge, J. J. (1978). Linguistic and psychometric factors in propositional reasoning. *Quarterly Journal of Experimental Psychology, 30,* 705–716.

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

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition.* Oxford, England: Basil Blackwell.

Staudenmayer, H. (1975). Understanding conditional reasoning with meaningful propositions. In R. J. Falmagne (Ed.), *Reasoning: Representation and process in children and adults.* Hillsdale, NJ: Erlbaum.

Staudenmayer, H., & Bourne, L. E. (1978). The nature of denied propositions in the conditional reasoning task: Interpretation and learning. In R. Revlin & R. E. Mayer (Eds.), *Human reasoning.* New York: Wiley.

Strawson, P. F. (1950). On referring. *Mind, 59,* 320–344.

Thompson, V. A. (1989). *Conditional reasoning: The necessary and sufficient conditions.* Unpublished master's thesis. University of Western Ontario,

Wason, P. C. (1983). Realism and rationality in the selection task. In J. St. B. T. Evans (Ed.), *Thinking and reasoning: Psychological approaches.* London: Routledge, Chapman & Hall.

Wason, P. C., & Johnson-Laird, P. N. (1972). *The psychology of reasoning: Structure and content.* London: Batsford.

Appendix

Table A1
Some Examples of the Relation Between Number of Models and Rated Difficulty in the Data of Braine, Reiser, and Rumain (1984)

Problem	Premise models	No. of models	Rating
	Problems with valid conclusions		
4			
o and z	o z		
? o		1	1.42
9			
if c or h, then p	c p, c, c p		
c	h p c h p		
? p	...	5	2.50
12			
if f, then l	f l, r l, f l r, f , f l r		
if r, then l	... ... r		
f or r			
?l		6	2.61
43			
l and, r or w	l r, l r z, l r z, l w z, l w r z		
if l and r, then z	l w ... l w r z ...		
if l and w, then z			
? z		7	3.86
58			
b or z	b , ¬z, b ¬z, b ¬r, b ¬z ¬r		
not z	z ¬b r		
not both b and r	¬b ¬r		
? not r		8	5.33
60			
l or w	l, l ¬e, l ¬e, w ¬e, w l ¬e, e , w l ¬e o		
if l then not e	w ... w l ¬e ... o		
if w, then not e			
e or o			
? o		10	5.80
	Problems with conclusions inconsistent with the premises		
2			
not m	¬m		
? m		1	1.18
21			
if e, then not k	e ¬k, e, e ¬k		
e	...		
? k		3	2.04

Table A1 (*continued*)

Problem	Premise models	No. of models	Rating
	Problems with conclusions inconsistent with the premises (*continued*)		
28			
if a and m, then not s	a m ¬s, a, a m ¬s, m, a m ¬s		
a	...		
m			
? s		5	2.79
20			
if e, then not v	e ¬v, o ¬v, e ¬v o, e , e ¬v o		
if o, then not v	... ... o		
e or o			
? v		6	3.17
44			
b and, t or z	b t, b t n, b z n, b z n, b z t		
if b and t, then n	b z ... b z t n ...		
if b and z, then n			
? not n		7	3.91
61			
e or x	e , e ¬h , e ¬h, x ¬h, e ¬h x , h , e ¬h x t		
if e, then not h	x ... x e ¬h ... t		
if x, then not h			
h or t			
? not t		10	6.00

Note. Each entry shows the problem number in Braine et al.'s appendix, the premises, the models required to interpret the premises, the number of models, and the difficulty rating of the problem.

Received January 18, 1991
Revision received May 6, 1991
Accepted July 22, 1991 ■