when it was wrong. To an A and an I premise, 79% thought it was correct and 73% thought it was correct; when it was correct. To an A and an E premise, 74% thought it was correct and 75% thought it was correct; when it was correct. To an A and an O premise, 36% thought it was correct and 39% thought it was correct; when it was correct and 39% thought it was correct.

This seems to me to be conclusive evidence that in both sets of these subjects were simply matching the logical form of one of the premises they had not engaged at all with the logical task. In my own studies [e.g., Wetherick & Gilhooly 1990] some subjects behave in this way and some are not so engaged with the logic, the crucial indicator is the degree of success achieved on syllogisms having a valid conclusion that is not of the same logical form as either premise. J-L & B unhesitatingly accept their subjects' performance when it is correct and ignore it when it is wrong. The set of H syllogisms in which the valid conclusion is of the same logical form as one of the premises happens to comprise the 10 said to require one model model and the 4 (with an A and an O premise) said to require two. Thirty syllogisms are said to require three mental models; that is, the 5 already mentioned (to which no correct conclusions were proposed) and having an A and an O premise and a valid O conclusion (to which 18 correct conclusions were proposed; see 160). They all have valid conclusions of a logical form different from that of either premise, conclusions that cannot be obtained by matching, hence the apparent relationship between problem difficulty and number of models required. Neither the correctness of correct responses nor the incorrectness of incorrect responses has been shown to have anything at all to do with number of models.

I conclude that mental-model theory has not been shown to have any claim on our attention as a scientific theory explaining deduction. "Requires more models" seems simply to be an alternative way of saying "more difficult" that contributes nothing to our understanding of the nature of the difficulty.

The mind may also use rules with specific content. Deduction reports experiments in the main are areas of deductive reasoning, and their results corroborate mental-model theory and count against the psychological theory based on formal rules. A gratifying number of reviewers accept our argument but some have misunderstood it, and so we will clear their misapprehensions before we reply to our critics. Some reviewers take us to have said (unsoundly) that the model theory is better than any other possible theory (Braine, Fetzer, Wetherick). In fact, our claim was not that the theory is better than other existing psychological theories, not that it is God's truth. We did not even claim to have excluded all possible theories based on formal rules after all, a computer program implementing a mental-model theory is a formal rule theory (as Bum Inder, and we ourselves point out, Deduction p. 213). A we certainly did not argue against rules per se (Andrews, Stenning & Oberlander, and Ter Meule). Comprehension and reasoning rely on syntactic rule semantic rules, rules for constructing models, and so on. If readers feel happier referring to them as "tactic inference rules" (Braine), so be it.

To those who believe there may be a superior theory based on formal rules or some tertium quid (Buach), we only say: we agree, but until someone formulates such a theory, the point is not a more thrilling (to us) than the claim that any scientific theory may be superseded. Those who believe that we argued for the model theory over all other possible theories including formal or syntactic ones, we can only cite the following passages: "A number of data, of course, can pick out one theory again all others" (p. 194): "Mental models and formal rules do not depend on syntactic procedures. . . . This claim is true for the computer programs modelling both sorts of theory" (p. 213).

One other misunderstanding arose about the formal rule theories. They postulate that difficulty reflects the number of steps in a deduction, and the relative availability, or ease of use, of the relevant rules. Commentators take us to task over these measures: they do not reflect differences between direct and indirect proofs (Fetzer) and a better measure is the number of embedded assumptions (Crawford). The measures, however, were not of our devising, but proposed by Braine, Rips, and other rule theorists (see pp. 29–31), who will doubtless appreciate these criticisms. Even so, the new measures fail, for example, to save the formal theories of spatial inference.

In our account of formal rules, we concentrated on so-called natural deduction systems. We did not describe Hintikka's (1955) model-set method, or the related "tableau" method of Beth (1955) and Smullyan (1968), though we did refer to them as formalizing the search for counter examples (p. 16; price Fetzer). The reason for our focus is that psychologists have not adopted tableau methods. Andrews regards mental-model building as a kind of tableau development; and Grandy emphasizes that disjunctions increase difficulty for both models and tableau. There is a resemblance, but Andrews points out there are differences—particularly in our use of implicit representations. Fetzer even constructs tableau proofs with lengths that do not predict difficulty—human reasoners are evidently not using these particular tableau rules.
So much for the main misunderstandings of the book. Our present goal is to help readers – and reviewers grateful of Reverend Smith's advice – to make up their minds about it. Its argument still stands, and, thanks to the commentators, it can be clarified and strengthened. We will concentrate on the controversial issues, but to those who agreed with us too, we are grateful for their encouraging words. We will try to deal with those notions that people found difficult to understand because our exposition was unclear. We begin with inductive competence and the algorithmic theory of performance. We then discuss each of the separate areas of deduction: propositional, relational, quantified, and informal reasoning. Finally, we consider rationality, the theory's extensions and deficiencies, and the possibility of combining it with formal rules.

R1. The nature of logical competence. Have we overestimated the importance of logical ability? Some reviewers think so (Fisher, Galetti & Komatsu, and Luchins & Luchins). Unfortunately, apart from the validation of certain intelligence tests, there is no objective data. We did concede that other forms of thought, such as creativity, are more important, but they are also much harder to understand. We believe that a world without science, technology, law, and social conventions (p. 3). And if the controversy about it is not resolvable, what hope is there for cognitive science?

In characterizing deductive competence, we argued that people are rational in principle, but that they transcend logic because their conclusions are parsimonious, maintain semantic information, and establish propositions not explicitly asserted in the premises. If no conclusion meets these constraints, then people say nothing follows. Hence, contrary to Fetzer's suggestion, we do not believe that formal systems are normative of human competence: they allow infinitely many different valid conclusions from any set of premises, including conclusions that people would never draw. Most commentators appear to accept our account of competence, but three reject it. Their objections do not seem to be decisive.

Modus ponens throws away semantic information, and so Cohen concludes that we are wrong, either to assume that it is a respectable deduction or to hold that deduction ought to maintain semantic information. In fact, we deal with modus ponens (p. 26). A special case of parsimony is not to draw a conclusion that asserts something that has just been asserted. Hence, given the premises:

If James is at school then Agnes is at work.
James is at school.

the conclusion:

James is at school and Agnes is at work.

is valid, but violates this principle of parsimony, because it repeats the categorical premise. This information can be taken for granted and, as Gaze (1975) argued, there is no need to state the obvious. It is therefore not a counterexample to our account: there is no need to reassert the categorical premise.

Cohen also claims that our studies presuppose rational competence, and that our argument would collapse if this assumption had to be abandoned. We agree that one cannot use errors to corroborate a theory without knowing what counts as an error. We presuppose the semantic principle of validity: an argument is valid if the truth of its premises guarantees the truth of its conclusion (p. 8). We accordingly analyze the truth conditions of premises and conclusions to determine what is valid. On this basis, our experiments show that nearly everyone is likely to make logical errors. Our analyses of truth conditions may be flawed, and indeed Cohen challenges our account of conditionals (see below). Such challenges, however, are normal in scientific criticism. We grant that if the semantic principle of validity is wrong, then our argument collapses – it is doubtful whether any argument about any topic could be usefully pursued.

Luchins & Luchins argue that untrained individuals draw redundant and contradictory conclusions, and that we overestimate their ability. People do make mistakes, but competence must not be confused with performance. Following Chomsky (1965), we treat a specification of competence as an idealization, and we assign the task of explaining errors to the algorithmic theory: the algorithmic theory should explain the characteristics of human performance – where it breaks down and leads to error, where it runs smoothly, and how it is integrated with other mental abilities" (p. 17). Our subjects make many deductive errors, but they rarely violate the standards of parsimony and novelty. Their valid conclusions never throw semantic information away by introducing new disjunctive alternatives.

Savion characterizes the model theory as too idealistic, and questions whether the semantic principle of validity is part of human competence, because she has found it difficult to teach to college students. This claim is akin to arguing against Murr's (1952) theory of vision because it is difficult to teach to students. Murr postulates mental models as the end product of vision, and he too might be accused (at least by Savion) of presupposing an enormous core of competence. Yet, in both his case and ours, it is possible to construct computer programs implementing the theory. The moral is that Savion overlooks the distinction between making a valid deduction and knowing what one is doing (p. 19, p. 147). Rats can make transitive inferences, according to Davis, but like Savion's students they probably do not know what they are doing.

R2. Images, mental models and logical models. According to the algorithmic theory, reasons construct models of premises and search for alternative models that are counterexamples to their conclusions. We now consider four aspects of the theory: the relations between mental models and images, the relations between mental models and their counterparts in logic (the model structures of Tarski), the search for counterexamples, and the claim that the theory is a set of separate microtheories.

The mental-model theory has its origin in the introspections of some subjects carrying out syllogistic inferences: their reports of using 'images' fitted Craik's (1943) theory of thinking. Subjects sometimes report that they reason 'verbally,' but they never report using formal rules of inference. We do not reject introspective evidence (pace Braine), but likewise we do not regard it as sufficient to eliminate formal rule theories. We were remiss in not pointing out the theory's kinship to the ideas of Max Wertheimer, one of the founders of Gestalt psy-
chology, and we thank Luchins & Luchins for pointing out the
semblances.

Images are a special case of models (p. 39; see also
Johnson-Laird 1983, Ch. 7; Pace Stenning & Oberlander
and ter Meulen). We hold this position for three reasons.
First, there are no detectable differences in performance
between those who claim to use images and those who do
not. Second, manipulations of imagery have no reliable
effects on deduction (p. 140, Johnson-Laird et al.
1989; Newstead et al. 1982; Richardson 1987), although
MacLennan points out that it is difficult to design decisive
studies. Third, our evidence suggests that models contain
tokens representing negation (p. 130). Negation cannot be
visualized: Signs representing negation can be visualized,
but the essential work is done not by the sign but by the
procedure for interpreting it. Similar annotations can be
invoked in responding to Over, who argues for the need
to distinguish between assumptions and beliefs (of varying
degrees of strength). Models represent propositional
content and contain separate annotations representing
epistemic attitude. MacLennan suggests that the annotation
representing negation should be treated as an intention
("Denial that") toward an image (see also Inder), and
that a disjunction of images works as well as a disjunction
of models. The snap is that models can represent many
assertions that cannot be readily visualized, for example,
"all men are mortal." MacLennan, however, argues that
the distinction between images and abstract models is one
degree rather than kind. We accept his point that
presence in consciousness is a matter of degree for both
models and images (see Yates 1985). The content of a
model may be available to consciousness, but the process
of inference and the format of mental representations are
never fully accessible (pace Braine and Wetherick). If
they were, then introspection alone would resolve most
controversies about mental representation.

What is the relation between mental models and
models in the logical sense that Tarski made famous? The
question is raised by Barwise, Bundy, Inder, and ter
Meulen. Barwise and Bundy emphasize that human be-
ings cannot construct models in the Tarskian sense. We
agree; there are too many possible models (p. 16, p. 36).
And Bundy's cautionary tale about his reasoning program
is also valid: programs must exploit role-like maneuvers.
Yet, he too respects the distinction in logic between proof
theory and model theory. He takes the mental-model
theory to imply that a corresponding computer program
would give meanings to computational states. He pro-
poses cogent arguments against this consequence; we
accept them. The application of the adjective "semantic"
is an existing computer program an oxymoron. Models
in programs, however, should not be confused with
models in minds. Mental models can genuinely represent
the world and the meaning of discourse because of their
causal relations to the world (p. 213; Johnson-Laird 1983,
p. 399 et seq.). Hence, as Inder says, they can reasonably
be claimed to be semantic.

We wrote that a mental model functions like a represen-
tative sample from the set of possible Tarskian models
of a statement (p. 36). This claim is wrong once negation
is introduced into models: as Inder points out, one model
then represents many states. Barwise proposes a subtle
but better way to construe the psychological theory: a
mental model represents a class of Tarskian models. This
proposal makes no difference to the empirical content,
the theory, but it clarifies its formulation. Consider a
example akin to Barwise's:

The triangle is on the right of the star.
The star is on the right of the line.
The line is on the right of the circle.

It follows that:
The triangle is on the right of the circle.
but it does not follow that:
The line is on the right of the circle.

As Barwise says, reasoners searching for a Tarskian mod-
le, that refutes the second conclusion may succeed, and
know that the inference is invalid, but those searching for
a Tarskian model that refutes the first conclusion will, of
course, never succeed, and so must ultimately abandon
the search. They will be right to do so, but they cannot
know that they are right; they cannot know that the
inference is valid. According to the model theory, how-
ever, the premises yield the model:

This represents an infinite class of possible Tarskian
models in which distances, sizes, shapes, and so on, are
all varying. There are only finitely many rearrangements
of the objects in a model of this sort, and so they can be
examined exhaustively (as in our program for spatial
inferences), and none of them refutes the valid conclusion.
The advantage of mental models (representing infinite
classes of situations) is that only a finite number need to be
explored to validate deductions of this sort.

The search for counterexamples is at the heart of the
model theory, but Polk remarks that it appears to under-
lie few predictions. The reason for the apparent anomaly
is simple: if the search is properly carried out, it yields a
correct response. Subjects often fail to carry it out proper-
ly, and so our predictions emphasize the increasing
difficulty of deduction as the number of models increases,
and the likelihood of errors based on a proper subset of
the possible models. Hence, the predictions are found on
the search for counterexamples.

Bara points out that even adults often fail to search for
counterexamples, and that it took years for Popper's
"falsification" criterion for demarcating hypotheses to
prevail over verificationism. The need to search for
counterexamples is not obvious. In fact, it is not a self-
conscious principle for the typical thinker. The best
evidence for its existence is that individuals respond "no
valid conclusion" reliably better than chance to premises
that do not validly yield informative conclusions (see also
Oakhill & Johnson-Laird 1985a). And Barwise makes a
point that we did not exploit: people can know that a
conclusion does not follow validly, but this knowledge
cannot be accounted for by formal rules.

Fetzer believes that there is a crucial equivocation in
our theory: people might think there are no counterex-
amples when in fact there are, and so they might believe
that arguments are valid when they are invalid, or that
they are invalid when they are valid. Exactly! Individuals
make both sorts of error: they fail to find a counterexample,
and they fail to see that a set of models supports a common
conclusion. But there is no equivocation: the theory
accounts for both sorts of error and for valid arguments, namely, when no counterexamples exist in a domain with a finite number of visual models.

To accommodate a new logical term it is only necessary to describe its meaning; that is, its contribution to the construction of models. The standard inference procedure can then take over (p. 272). Hence, the development of the theory has been piecemeal. Polk, while commending the completeness and accuracy of the theory, suggests that it is really a set of microtheories based on a common framework. He lists a set of assumptions, which he claims are not part of the core theory but of particular microtheories. In fact, all of the assumptions that Polk lists are part of the core theory:

1. The greater the number of its atomic propositions, the harder an inference; as the number increases so does the size of the models.
2. Individuals reason only about those items that are explicit in their models. This assumption applies to any form of reasoning (see Legrenzi & Sonino).
3. Reasoners consider multiple models of premises even if they all support the same conclusion (more on this point below).
4. The need to integrate premises by bringing their referents in common into contiguity applies to any form of reasoning.
5. "Negative" deductions, which call for the deduction of an inconsistency between one model and another, are always harder than affirmative deductions.
6. It is easier to reason from the hypothesis that an assertion is true than the hypothesis that it is false, because it takes work to construct the complement of the set of models representing a hypothesis.

Nevertheless, Polk is right about the origins of the theory. We were guided by the general framework in formulating a semantics for each logical domain, and these accounts are microtheories, that is, we can formulate them in many ways, which are independent of the inferential mechanism (the search for counterexamples). Otherwise, the components of the theory that appear to be local to particular domains turn out to be general.

R3. Propositional reasoning. An exclusive disjunction, such as "Either there is a circle or else there is a triangle, but not both," has the following initial models:

\[ [O] \quad \neg [\Delta] \]

where the brackets indicate that a proposition has been exhaustively represented. The procedure for fleshing out models works as follows: when a proposition has been exhaustively represented in one or more models, it adds its negation to any other models. The procedure adds that there is not a triangle to the first of the two models above because triangles are exhausted in the second model. The final result of fleshing out the models is:

\[ [O] \quad [\neg \Delta] \quad \neg [O] \quad [\Delta] \]

where "-" denotes negation. These two models correspond to those rows in the disjunction's truth table that are true. In general, the number of explicit models for a propositional deduction equals the number of rows that are true in a truth table of all the premises. Exhaustion is therefore a device that allows the inferential system to represent certain information implicitly—it can be made explicit but at the cost of fleshing out the models. Andrews grasps the basic idea but wonders how it applies to complex propositions. Hodges gives up on it and, it seems, on the model theory as a whole; perhaps the problems of psychology are too murky for him after the clarity of logic, though we agree with him that cognitive scientists from different disciplines should talk to one another. Andrews asks how exhaustion applies to the representation of the sentence:

Either there is a circle or a triangle, or a triangle and a square, but not both.

This is a reasonable question from a logical point of view, but it misses the psychological point. The answer is that exhaustion could be used recursively. The procedure represents the main connective (A or B, but not both) first:

\[ [A] \quad [B] \]

It then represents proposition A (there is a circle or a triangle) which we will assume to be an exclusive disjunction:

\[ [O] \quad [\Delta] \]

and proposition B (there is a triangle and a square):

\[ [\Delta] \quad [\Box] \]

Finally, it substitutes these models in place of A and B in the initial model:

\[ [O] \quad [\neg \Delta] \quad \neg [O] \quad [\Delta] \]

Fleshing out can be accomplished in a similarly recursive way. The overall models become:

\[ A \rightarrow B \]
\[ \neg A \quad B \]

The model corresponding to B is already fleshed out, and the model corresponding to A fleshes out as:

\[ O \neg \Delta \]
\[ \neg O \quad \Delta \]

A and B (and their complements) yield the final explicit models:

\[ A \rightarrow B \text{ yields: } O \neg \Delta \quad \Box \]
\[ \neg O \quad \Delta \quad \neg \Box \]

and \[ \neg A \quad B \text{ yields: } O \quad \Delta \quad \neg \Box \]

Even here, recursive exhaustion is too complicated to be psychologically plausible. Our study of "double disjunctions" shows that individuals have difficulty in representing three distinct models (p. 56).

R4. Conditionals. Our theory of indicative conditionals combines Gries's (1975) conversational implicatures with implicit representations. The antecedent and consequent are represented explicitly in one model, and there is an alternative wholly implicit model. These initial models yield judgments corresponding to a "defective" truth.
table, difficulty with *modus tollens*, and difficulty with Wason's (1966) selection task. If the models are fleshed out, they yield a truth-functional interpretation. Two reviewers express qualms about this account. Over argues that an indicative conditional tends to be asserted and accepted when its consequent seems highly probable given its antecedent. We give no account of this reasoning, because we doubt it. At the time of writing, for example, we believe that, given that Bush is nominated, Clinton is highly likely to win, but we would not accept, assert, or consider to be true, the following conditional:

If Bush is nominated, Clinton will win

though we would accept:

If Bush is nominated, then Clinton is likely to win.

Hence, our doubts over Over's claim.

Cohen cites the following strange inference (see also Cooper 1968), which he regards as problematic for our analysis:

If John's automobile is a Mini, John is poor, and, if John's automobile is a Rolls, John is rich.

Therefore, Either, if John's automobile is a Mini, John is rich (i.e.) or, if John's automobile is a Rolls, John is poor (sic).

The inference is valid, though it seems not to be. However, it throws semantic information away with a vengeance. Even assuming that John cannot be both rich and poor, the conclusion has eleven models. We claim that hardly anyone can mentally envisage these explicit models for a *disjunction* of the two constituent conditionals. Taken individually, the conditionals conflict with those in the premises, and so the inference seems invalid.

We reported a study in which subjects balked at *modus ponens* (p. 81 et seq. Byrne 1989). When we gave them such premises as:

If Paul goes fishing, then Paul has a fish supper
If Paul catches some fish, then Paul has a fish supper
Paul goes fishing

they do not draw the conclusion:

Paul has a fish supper.

Contrary to the claim that formal rules are automatically applied to any assertions of the appropriate logical form. Several reviewers objected to our referring to the "suppression" of *modus ponens*, and suggested instead that subjects reject one of the premises (Bach), or take the second conditional to render the first one uncertain (Over) or false (Savion; see also Politzer & Braine 1991, and for a response Byrne 1991). Our claim, like Grandy's, is that subjects interpret the two conditionals as equivalent to "If Paul goes fishing and catches some fish, then Paul has a fish supper." They produce this sort of paraphrase of the two conditionals and they do not say that one conditional renders the other false (Byrne & Johnson-Laird 1992). In the light of the reviewers' claims, Byrne and Handley carried out an experiment in which the subjects explicitly judged the truth values of the conditionals: they tended to judge that both conditionals were true, especially after they had carried out the inferential task. Fillenbaum's judicious discussion leads him to a conclusion with which we concur: the interpretative component is critical. Rule theories need to explain recovery of logical form, and the model theory need to explain how background knowledge produces cues corresponding to those of the paraphrase.

The model theory yields an explanation of performance in Wason's selection task (pp. 75-81). It is the only account that purports to explain the effects of all the variables including abstract and realistic materials. Evans regards it as perfunctory, but does not explain why. Green correctly points out that the critical issue concerns the factors cue subjects to flesh out their initial models. I report a study that confirmed the core of the model theory, but some subjects who envisaged the critical counterexample failed to select it. There are possible reasons for such a failure, such as the liability of the identification. We take comfort in the overall correlati between identifying counterexamples and making correct selections.

Mankelow points out a real difficulty for our account, but in fact it is a problem for any theory. Insightful subjects should select all four cards if they have made an equivalence interpretation of the conditional. Perhaps they would select all four cards if the experimental procedure did not imply that this selection was somehow redundant. Mankelow also argues that we need to distinguish between evaluations of truth value and evaluations of violations of rules, and that we need a procedure to convert (invalidly) the conditional to a "if only if" form in order to explain "p and q selections. In fact, we do propose a verbal conversion of "if p then q" to "q only if p" if we argue that the interpretation of the rule in context yields the "only if" models. However, we entirely agree that there is a difference between judging truth value and judging violations. The theory's formulation in the latter case should read: select those cards that have bearing on compliance or violation of the rule. We also accept that preferences and "point of view" matter (C Johnson-Laird & Byrne 1992). Our only doubt is whether preferences are based on immediately available "set" theories. What, for example, is the utility to us of Mankelow's commentary? Certainly we would give it positive value, but to obtain a utility in the classic decision-theoretic sense would be difficult, and would itself require us to reason.

Pallard remarks that the rules yielding the most insight into the selection task are those for which counterexamples are known to the subjects. Hence, he urges us to adopt a "nonological" model of the conditional that represents explicit counterexamples. We accept this recommendation, but with one qualification: The counterexample is not part of the models for the conditional but is represented as impossible (given a true rule) or as implies (given a deontic rule). Indeed, we hinted at this very idea (p. 80). To represent the conditional as embracing the counterexample would be a little too "nonological".

R5. Syllogisms. Commentators raised more questions about our account of syllogistic inference than about any other topic. The most serious question was: Is there an objective procedure for generating the models? We have implemented the theory in a computer program; it differs in detail from the theory in Chapter 5, but it yields the same general predictions. The principle for combining models is, as Garnham surmised, to combine those indi
which supports only the conclusion: “Some of the C are A.” In general, where only one model is needed to formulate a valid conclusion, the task should be harder if the conclusion is consistent with a further model that falsifies another conclusion. Hence, as in this example, AA premises yielding valid A conclusions should be harder than IA or AI premises yielding valid I conclusions. (We use the abbreviations: A for assertions of the form “All X are Y,” I for “Some X are Y,” E for “No X are Y,” and O for “Some X are not Y.”) The evidence appears to bear out this prediction (Table 6.1). This explanation should help Newstead to explain the effects of belief upon syllogistic inference. It also lays to rest one of Polk’s worries. He points out that reasoners are affected by multiple spatial models, even if they support the same conclusion, but not similarly affected by multiple syllogistic models. The same difficulties occur in both domains.

Newstead and Pollard both wonder whether an O premise, “Some of the A are not B,” may pragmatically imply the truth of the I premise, “Some of the A are B.” One small class of conclusions can be explained in this way. For example, premises of the form:

- Some of the A are not B
- All of the B are C

elected the following percentages of responses (over four experiments with correlated results):

- Some of the A are not C (52%)  
- Some of the C are not A (7%)  
- No valid conclusion (35%)

These are the responses predicted by the model theory. The remaining responses were almost all:

- Some of the A are C (15%)

which may reflect the pragmatic interpretation of the O premise.

Stenner & Oberlander have shown how a new method of using Euler circles is equivalent to the model theory. We applaud their ingenuity. They claim that the method is more constrained than mental models, but it is not clear why. Nor is it clear why these authors believe that we rule out graphical methods. Images are a special case of models, and presumably they are graphical. What we do maintain, however, is that traditional Euler circles cannot represent multiply quantified relations, and so they are unlikely to be used by logically untrained individuals, who move freely from singly to multiply quantified relations (pp. 134–35).

Wetherick rejects our account of syllogistic reasoning and argues that some subjects are prey to an “atmosphere” effect in which they match their conclusions to the mood of one of the premises, whereas other subjects “engage with the logical task.” Some subjects may sometimes draw a conclusion because it matches the mood of a premise, but matching is implausible as a general account of syllogistic reasoning. First, there is a simple alternative explanation: the initial model of any conventional syllogism yields a conclusion matching the mood of at least one premise. Matching, however, cannot explain why sub-
jects ever respond with conclusions that emerge from subsequent models or why they respond, "there is no valid conclusion." Second, a striking failure to demonstrate matching occurred in a study of "only" as a quantifier. When syllogisms contain "only," subjects were most reluctant to draw a matching conclusion, preferring instead the quantifier, "all" (p. 129). Third, matching fails to explain performance with multiply quantified premises (p. 140). Wetherick accuses us of overlooking an alternative explanation for our results. The truth is that we have not overlooked it, but eliminated it. Indeed, matching seems to be a more accurate account of its own origins than of reasoning.

R6. Multiple quantifiers. Existential quantifiers should be more difficult than universal quantifiers, Grandy argues, just as conjunction is more difficult than conjunction. The hypothesis is aesthetically pleasing; the facts are ugly. To best syllogism contains an existential (Some of the A are B), All the B are C, and the easiest double quantified problems, as he acknowledges, include those with existentials. Conversely, some difficult deductions do not contain existentials. He points out that proofs can be difficult because of the interactions among the rules for quantifiers. We leave to rule theorists the task of determining whether these interactions could in principle explain our results with multiply quantified deductions. We doubt it. Our untrained subjects generate conclusions in a minute or so; we have yet to see logicians derive new conclusions not just prove given conclusions in a comparable time. Either our subjects have a remarkable tacit ability at logical derivations (which is, alas, not available in the logic classroom), or, as we believe, they are reasoning by other means.

Crawford argues for an intrinsic logical difficulty of \( \exists \exists \) quantification, that is, as when the assertion: "All musicians are related to some authors" is interpreted as all musicians are related to some author or other. The reason is that assertions of this form may call for an unbounded set of individuals (see also, pp. 178–80). Once again, the theory is beautiful, but not the facts. We found no differences in reasoning with assertions of the \( \exists \) form and the \( \exists \exists \) form (pp. 142–45). Crawford suggests that deductions that are computationally tractable but difficult for people may yield evidence about human reasoning algorithms. We agree, but we note that people often work with small-scale problems in intractable domains, for example, they can deduce their own parsimonious conclusions from propositional premises.

R7. Everyday informal reasoning. Not only is most reasoning utterly unlike syllogistic or propositional reasoning, claims Fisher, but it is not even deductive (or inductive). He and many of his colleagues in the movement for "informal logic" believe that logic has little application to the analysis of everyday arguments (see, e.g., Toulmin 1958). Galatalt & Komatsu suggest, however, that the model theory may be a fruitful way to explain everyday arguments, and this approach looks promising (see pp. 205–6, and Morton 1988). The model theory shows how to combine valid deduction and the "nonmonotonic" reasoning that occurs in undoing arbitrary or default assumptions (pp. 180–83). And a major implication of the theory is that people are inferential satisfiers (cf. Simon 1991), that is, once they find a model that fits their beliefs, they do not tend to search for alternatives (p. 126). The negative tendency seems to be the cognitive cause of many disasters, for example, the operators at Three Mile Island thought that the high temperature of a relief valve arose from a leak and overlooked the possibility that the valve was stuck open. The theory makes the same prediction about everyday arguments: people draw conclusions that are true in some models of the premises and often overlook alternative models. Hence, even in daily life, counterexample is likely to devastate an argument effects are explainable only in terms of an underlying deductive competence (pace Fisher, and Luchins & Luchins).

Our theory assumes that general knowledge is used in constructing models, and so it dovetails with Tversky and Kahneman's (1973) "availability" heuristic (p. 206, p. 241). Any available knowledge can be embodied in model, and the process of comprehension makes some items of knowledge more available than others. The account finesse the mechanism underlying availability (as pointed out by Green, Loder, Legrenzi & Souina, and Stevenson). Inder remarks that we are assuming that the brain comes fitted with "mental-model accelerators," and we worry as not to push too much of the explanatory bone into cognitive architecture. We take the point, but retrieval of relevant knowledge is the only major problem for which we have presupposed a solution. Both Green and Stevenson suggest that expertise in a domain may lead to the development of content-specific rules, and Evans believes that such rules might even be implemented in a connectionist network (cf. also MacLean).

The evidence about implicit reasoning in comprehension suggests that people do construct models (see, e.g., Carnahan 1987), and the general knowledge used in their construction could be represented in content-specific rules (Newell 1990; see also multiple book review, BR, 15(3) 1992). So far, however, no evidence has identified how knowledge is mentally represented — it could be in the form of rules, assertions, models, or networks. For nonmonotonic reasoning, Chater suggests that the most plausible conjecture is that reasoning by searching for counterexamples is inapropriate because they always exist. We see no reason why the construction of models needs to be constrained by a logic: it is constrained by available knowledge. We likewise see no difficulty in basing nonmonotonic reasoning on a search for counterexamples. Consider Chater's own example: You infer from the sound of purring that your cat is trapped in the cellar, but you overlook the conclusion if you catch sight of it in the garden. Yet withdraw the inference because of a counterexample. And isn't this very example a counterexample to the claim that counterexamples are an inappropriate method of nonmonotonic reasoning? And so shouldn't the claim be withdrawn? Nonmonotonically.

Oakland argues that the model theory is unlikely to yield a tractable account of nonmonotonic reasoning. Perhaps not. But nothing implies that people are using a tractable procedure. Our experiments on everyday inference (p. 205) suggest that they bring to mind an available scenario. They think of some alternatives, but they overlook many possibilities. They may be using an intractable algorithm that is defeated by the magnitude of the task.
Response/Johnson-Laird & Byrne: Deduction

The model theory accounts for the undoing of arbitrary or default assumptions, but it does not explain another sort of nonmonotonic reasoning. When you draw a conclusion that clashes with your beliefs, you have to reconcile the discrepancy. You may revise your inference, your premises, or your beliefs. You search, as Bach emphasizes, for the best explanation (Harman 1986). How do you do so if largely unknown, but you may create new ideas. If we said that we had solved this problem (and Cluster takes us to have said so), then we were mistaken. It remains deeply mysterious.

87. Rationality versus relativism. We rejected relativism and defended universal rationality founded, not on formal rules of inference, but on the semantic principle of validity. Engel largely agrees with us, but argues for two sorts of rationality: rationality of purpose, and rationality of process (see Evans, in press). We accept the distinction, though we worry about its psychological basis—how, for example, are the two sorts of rationality mentally embodied, and how do they interact? Engel suggests that mental models may play a part in rationality of purpose, and perhaps they provide an underlying common framework. He also defends the process of "reflective equilibrium" that enables people to bring their intuitions and normative principles into harmony. People may go through such a process, but our concern is their basis, not their normative principles.

Braine considers the occurrence of logical terms in all human languages as more naturally explained by a mental logic than by the principle of semantic validity. It is difficult to advance beyond the mere truism of intuitions. However, relativists argue that no account of deductive competence justifies a unique system of logical rules (pp. 205–9), and Braine proposes no solution to this problem.

Only one commentator hints at a defense of relativism. MacLennan suggests that we should allow for culturally permissible inconsistencies. We doubt that anyone would knowingly accept an inconsistency on the grounds that it is culturally acceptable—except perhaps as a Whorfianesque posture: "Do I contradict myself? Very well then I contradict myself. (I am large, I contain multitudes.)" Certain inconsistencies may be invisible to a culture, but do they remain so once some intrepid individual points them out? What matters is the recognition of inconsistencies, because consistency is a universal of logic (pace MacLennan). And so, we believe, is the semantic principle of validity.

89. Extensions of the mental-model theory. Numerous commentators who accept the core of our argument raise the possibility of extending the model theory to cope with other mental processes, including, as Bach and Calotti & Konatsu suggest, those for which formal rules would be implausible. Baron argues that a modest generalization of the theory can describe all goal-directed thinking: thinking is a search for possibilities, evidence, and goals. Search is biased towards positive evidence, he says, particularly because some people do not understand the importance of negative evidence. Another factor arises directly from models (p. 79): as Legrenzi & Sonino point out, individuals focus on what is explicit in their models and neglect what is only implicit in them. Tversky remarks that most scientific thought fits comfortably within the model framework, but he rightly points to the problem of how such models are initially created by scientists (see Johnson-Laird, 1992, for some thoughts on this issue). What is striking is the role that the manipulation of models appears to play in scientific innovations (see, e.g., Wise 1970). Bara argues that the model theory needs to be extended to the development of deductive ability in children. The beginnings of such an extension can be found in Johnson-Laird (1990b); intellectual growth is not the development of new mental operations (pace Piagetians) but the development of new concepts—a hypothesis urged by Carey (1991), Ezel (1991), and others. Verbal instruction is no substitute for the construction of models of the world.

Davis draws attention to the burgeoning literature on animal reasoning and asks whether the model theory has a role to play here. His results are suggestive, but the data are still hard to interpret. Mayr argues that the model theory can be used to account for the evolution of perceptual abilities (Mayr 1988), but it is unlikely to depend on a postulate of transitivity and the predicate calculus. Animals build up spatial models of their environment and combine models that have been separately learned. Because models probably owe their origins to the evolution of perceptual abilities (Mayr 1988), they are likely to play a role in the inferences of humans and nonhumans alike. What is unique to humans is natural language and, perhaps, the ability to reflect on their own performance.

90. The incompleteness of the research. We have almost completed our reply to the commentators' detailed queries, but we will consider some methodological criticisms before we review the current status of the theory. Ter Meulen has reservations about our experiments: she asserts that of manipulating the order of premises, and of using materials of a limited genre. Evans also claims that our studies, except for those on belief-bias, use materials with an arbitrary content. These claims are incorrect, though we could not describe all our experimental manipulations in a research monograph—a limitation for which readers should be profoundly grateful. We did, however, report one major effect of order of premises (Table 6.2; for other effects, e.g., Erlisch & Johnson-Laird 1982, Legrenzi et al. 1992), and we used realistic materials in our studies of the "suppression" of nandus ponsus (pp. 81–84, Byrne 1989), paraphrasing with conditionals (pp. 84–85; Byrne & Johnson-Laird 1992), and informal everyday inferences (pp. 205–6). The major studies of formal rule theories have used arbitrary materials, and for purposes of comparison we used them too.

As we wrote in Deduction, the model theory is incomplete (p. 213). It does not explain, for example, how the search for counterexamples is carried out; our experiments on the topic were not successful. Psychological experiments familiar with this problem, and where there are no grounds for theorizing, do not demand that a theory be specified to the last detail. Our colleagues in other disciplines, however, chide us about the theory's incompleteness:

What counts as a mental model?

What other forms of mental representation are there apart from models?
How does the theory apply to nonstandard quantifiers or to modal reasoning?
What drives the search for alternative models?
How do people know when no more models are needed?
Do two people have the same mental models if they make the same inferences?
We have tried to answer the first three questions (Johnson-Laird 1983), but we do not know the answers to the rest.

Some reviewers found our account incomplete in at least one way in which it is not (Bach, Falmagne, ter Meulen). They ask: How are models constructed from premises? We described a compositional semantics for connectives, relations, and quantifiers; and we showed how it is implemented in computer programs (see Ch. 9). We can provide more detail if anyone requires it. Fodor remarks that to account for all aspects of performance places too much of a demand on the theory: performance can be affected by motives, beliefs, ethics, ability, and other factors. A complete theory of the mind would aim to embrace such effects, and our theory aims to account for effects of beliefs and ability.

The question of individual differences drew comment from Bach and ter Meulen. Bach asks whether some ‘hard’ problems might be easy for some people. The answer is that the differences between moderately hard problems and easy ones can disappear for exceptional reasoners, but we have never observed reliable reversals in difficulty. The trends are remarkably robust, for example, we have never tested anyone who does not do better with one-model syllogisms than with multiple-model syllogisms. Model-based accounts of individual differences are now under way, especially as a result of Fulk’s studies and his development of software for fitting parameterized theories to individuals’ data (Fulk & Newell 1992). Why, Bach asks, are some people better at deduction than others? The answer, judging by the patterns of systematic error, is that they do not appear to be vast differences either in models or strategies. We know that a measure of the capacity of working memory accounts for part of the variance, and that a measure of the ability to perceive what is common to different drawings accounts for rather more (Bara et al. 1992). But we do not know how best to interpret these correlations, and we are far from a causal account of differences in ability.

Finally, some reviewers believe our theory is incomplete because it does not include informal rules. They urge us to combine models and rules. They particularly want the rule for modus ponens: Pollard asserts that without it individuals would have difficulty constructing models for conditionals. Falmagne argues that it is required for rapid automatic inferences relying on form, and Wetherick claims that no theory applying general principles to particular cases can avoid it. In fact, models for conditionals can be constructed without the rule for modus ponens: our computer programs construct them solely from the ‘truth conditions’ of the connective. People make modus ponens deductions rapidly, but it does not follow that they rely on a firm rule to do so — even if the content of the premises is abstract or based on nonce words. They could build models containing abstract or nonce items (pace Sagon).

And unified theorem-provers apply general principles to particular cases without using modus ponens: they rely instead on unification and the resolution rule (see pp. 27 for a brief account). With experience in a reasoning task, as Stevan shows, subjects may begin to construct formal rules for themselves (see p. 202; Galetti et al. 1986). But few individuals seem capable of forming rules that capture all the valid deductions they can make (as Victoria Shaw has found in an unpublished study).

The model theory proposes that variables occur in initial semantic representations of sentences (pp. 171–73), and that they are then instantiated in models as finite sets of tokens corresponding to individuals (pp. 177–90). Braine correctly claims that models do not contain variables, but he overlooks their occurrence in initial semantic representations (pp. 171–73). Ter Meulen understands that our programs use variables in this way but does not seem to realize that the programs implement our theory.

R11. Conclusions. No commentator proposes a new theory accounting for the phenomena of deductive reasoning. So: mental models or formal rules? Or both? A theory that combines both is probably irrefutably true, that is, an evidence could ever show it to be false. We therefore preferred to exercise parsimony and to reject the psychological theories based on formal rules. Braine claims the argument for mental models rests solely on parsimony; and Fodor claims that it rests solely on the effect of content. They overlook the real case. It is the experience of experimental evidence as a whole. There are psychological theories based on formal rules for relational reasoning but the empirical evidence counts against them. There are such theories for propositional reasoning, but they fail to explain differences in difficulty, systematic errors, or the effects of content. No such theories exist for syllogisms or multiply quantified deductions. In each of the domains, the model theory explains the phenomena.

Where do we go from here? Our immediate tasks are to develop better accounts of everyday reasoning, to explain how reasoning discovers new strategies in metareasoning (see Byrne & Handley 1992), to interpolate reasoning as a decision making, and to find out how diagrams impinge on reasoning — especially in the light of recent studies (Barwise & Etchemendy 1991; Bauer & Johnson-Laird 1993). We thank the commentators for helping us to strengthen the exposition of the theory, to clarify the relations between mental models and Tarskian models, and to pursue extensions of the theory into other realms thought.

References
Letters a and r appearing before authors' initial refer to target article or response respectively.
References/Johnson-Laird & Byrne: Deduction

[1990] Remembering, reasoning, we have inferred. What have we inferred? In: Cognitive limits. Their contribution to human deductive processes, eds. J. P. Cavanagh, J. L. Fowler & M. Gomolin. North Holland. [aPN-L]