Psychological Review
1995, Vol. 102, No. 3, 533-566

Copyright 1995 by the American Psychological Association, Inc.
0033-295X/95/$3.00

Deduction as Verbal Reasoning

Thad A. Polk
University of Pennsylvania

Allen Newell
Carnegie Mellon University

Most theories of deduction have assumed that linguistic processes transduce from language into an internal representation and back again, and that non-linguistic processes are central to deduction itself. In this article it is proposed that for deduction tasks for which the necessary information is provided verbally, the heart of deduction for untrained participants involves repeatedly reencoding the problem, a type of behavior referred to here as verbal reasoning. It is shown that model theory accounts of behavior on most deduction tasks are consistent with verbal reasoning and that verbal reasoning can account for detailed behavior in a single task; a computational model of syllogistic reasoning—VR—based on linguistic mechanisms is presented. VR models all of the standard phenomena, makes a number of accurate novel predictions, and fits the behavior of individual participants with an accuracy that rivals their own test–retest reliability.

Human beings are constantly faced with the need to reason, to infer something novel from available information. If John's friend tells him that she will be either at home or at work and he cannot reach her at work, John fully expects her to pick up the phone when he calls her house. John comes to this conclusion so quickly and easily that it hardly seems like reasoning. Yet John could reach it only by combining the information from at least two separate assumptions: (a) His friend is either at home or at work and (b) his friend is not at work (not to mention the assumptions that his friend told the truth, that the phones are working, and so forth). Neither assumption alone would lead John to the conviction that his friend is at home, so he must have put them together in some way. How did he do it? More specifically, what cognitive processes did he apply so effortlessly to reach that conclusion?

A particularly pure form of reasoning is deduction: determining logically valid consequences from information. A number of different hypotheses have been proposed concerning the nature of the cognitive processes that people use when reasoning deductively. Braine (1978) and Rips (1983), among others, have argued that human deduction can best be characterized as the application of logical inference rules. According to this view, the cognitive architecture implements a set of syntactic rules that can be used to draw inferences. For example, one such rule that is commonly proposed is *modus ponens:* If A and if A implies B, conclude B. Given a set of such rules, reasoning consists of encoding the problem into a form against which these rules can match (some kind of logical proposition), applying some of these rules to generate new propositions, and decoding the results into a form appropriate for the task (e.g., natural language). By assuming that the rule set is incorrect or incomplete in some way (e.g., missing a rule for *modus tollens:* If not B and if A implies B, conclude not A) and that some rules are more likely to be used than others, it is possible to account for correct performance on some problems while explaining errors on others. Problem difficulty is accounted for in terms of the number of rules that must be applied. Problems that require long "proofs" are predicted to be harder than those that require shorter ones.

Cheng and Holyoak (1985) have proposed a different type of rule theory to account for deduction. They hypothesized that participants sometimes solve deductive reasoning problems by using *pragmatic reasoning schemas.* Pragmatic reasoning schemas are "generalized sets of rules defined in relation to classes of goals" (Cheng & Holyoak, 1985, p. 391). The idea is that a reasoning problem might trigger the retrieval of a schema from memory and that inference rules are associated with this schema; these inference rules can then be used to make deductions. For example, a problem that involves determining who is allowed to drink alcoholic beverages is hypothesized to evoke the use of a permission schema. Associated with such a schema are a number of content-specific inference rules such as the following: If the action is to be taken, then the precondition must be satisfied. Errors can arise when these rules are incomplete or not strictly valid.

Johnson-Laird and his colleagues (Johnson-Laird, 1983; Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991) have argued against rule theories of deduction in favor of a men-

Thad A. Polk, Department of Psychology, University of Pennsylvania; Allen Newell, Departments of Computer Science and Psychology, Carnegie Mellon University.

This research was supported in part by a fellowship from the Eastman Kodak Company, by the James S. McDonnell Foundation under Grant 91-34, by the Office of Naval Research under Contract N00013-91-J-1527, and by the Avionics Laboratory, U.S. Air Force, under Contract F33615-90-C-1465. This article formed the basis of an unpublished doctoral dissertation (Polk, 1992). A summary of the argument that the mental model theory accounts of behavior on most deductive reasoning tasks correspond to verbal reasoning explanations can be found in Polk (1993).

We gratefully acknowledge Philip Johnson-Laird, members of the Carnegie Mellon University Soar group (especially Richard Lewis and Richard Young), and Graeme Halford for helpful comments on this research and article.

Correspondence concerning this article should be addressed to Thad A. Polk, Department of Psychology, University of Pennsylvania, 3815 Walnut Street, Philadelphia, Pennsylvania 19104-6196. Electronic mail may be sent via Internet to polk@psych.upenn.edu.

tal model theory. This theory proposes that deductive reasoning consists of three stages: comprehension, description, and validation. Comprehension corresponds to encoding the problem into an internal representation. Unlike rule theories, this representation is assumed to take the form of a mental model rather than a set of logical propositions or a schema. A mental model represents a scenario consistent with the problem statement. For example, if the problem states that all dogs have four legs, the corresponding mental model would contain a set of tokens corresponding to dogs, and all of them would be represented as having four legs. In contrast, rule theories would posit a single logical proposition containing a universal quantifier [e.g., $\forall x Dog(x) \rightarrow FourLegs(x)$]. The second stage of deduction, according to mental model theory, involves generating a putative conclusion. This stage requires formulating a description of some aspect of the mental model that was not explicit in the problem statement. Finally, in the validation stage, participants are hypothesized to search for alternative mental models that falsify the conclusion just generated in an attempt to ensure its validity. Mental model theory accounts for errors in terms of a failure to consider falsifying models. Problems that are consistent with more models are thus predicted to be harder than those that are consistent with fewer.

Despite their differences, these approaches have all been based on what we call a *transduction paradigm:* Participants encode the problem into an internal representation (be it logical propositions, schemas, or mental models), reason using operations that are specifically devoted to inference (applying formal rules or searching for alternative models), and decode or describe the result to produce a response. The only role played by the operations of encoding and decoding is transduction: translating the problem into a form that the reasoning-specific mechanisms can operate on and then translating back the results. Although these transduction operations are sometimes assumed to lead to errors in reasoning (e.g., through the construction of invalid encodings; see Henle, 1962), they are assumed to play only a peripheral role in reasoning. The heart of deduction, according to these theories, is in the application of reasoning-specific processes. As Johnson-Laird and Byrne (1991) have stated, for example, "only in the third stage [searching for alternative models] is any essential deductive work carried out: the first two stages are merely normal processes of comprehension and description" (p. 36).

We have come to believe that the linguistic processes of encoding and reencoding actually play a central role in deduction itself and that processes that are devoted exclusively to reasoning (applying inference rules and searching for alternative models) play a smaller role in comparison. We propose that the behavior of most untrained participants on standard deduction tasks can best be characterized as *verbal reasoning,* that is, the deployment of linguistic processes according to and so as to satisfy the demands of a reasoning task.

Verbal Reasoning

Newell (1990) argued that cognitive science can and should work toward unified theories of cognition, that is, "theories that gain their power by positing a single system of mechanisms that operate together to produce the full range of human cognition" (p. 1). He proposed a computer system called Soar as an exemplar unified theory of cognition and, with John Laird and Paul Rosenbloom, began an effort to apply Soar in as many domains as possible (Laird, Newell, & Rosenbloom, 1987; Newell, 1990).

The idea of verbal reasoning arose out of one such attempt, namely to model human syllogistic reasoning behavior within the Soar architecture (Newell, 1990; Polk, 1992; Polk & Newell, 1988; Polk, Newell, & Lewis, 1989). Implementing a theory based on the transduction paradigm is unnatural within Soar, because Soar does not contain any processes or mechanisms that are devoted to reasoning (or to encoding and decoding for that matter). Indeed, the mechanisms in Soar (e.g., searching in problem spaces, subgoaling to other problem spaces, and learning by summarizing previous problem-solving experience in new production rules) were specifically designed to be general-purpose rather than task-specific mechanisms. We were thus forced to ask whether the processes used in reasoning could be acquired. And although it was clear that people can and do acquire the verbal operations of encoding from and decoding into language, it was much less clear whether people acquire processes devoted exclusively to reasoning without any training. Consequently, we began to explore the idea that verbal processes might be more central to the deduction of naive participants than are sophisticated reasoning-specific mechanisms. Given that all standard tests of deduction are presented verbally and require a verbal conclusion, the task clearly calls for linguistic processes to encode the premises and to produce a response. And although participants' linguistic processes are extremely sophisticated, we assume that their reasoning-specific processes are much less well developed (or perhaps, in some cases, even missing entirely). If one is faced with a deductive reasoning task, then, a natural approach would be to attempt to adapt one's well-developed linguistic skills to the demands of the task. For example, if a conclusion is not obvious after the initial encoding of the problem statement, then participants could use their linguistic skills to reencode the problem repeatedly in an attempt to elaborate their internal representation until it makes a conclusion explicit (or until they give up). Note that if this explanation of participants' behavior is correct, the operations of encoding and reencoding are not just playing the role of transduction; they are at the very heart of reasoning itself. It is this kind of behavior that we refer to as verbal reasoning.

Although verbal reasoning is a natural strategy for solving deductive reasoning problems, it can often lead to errors. The problem is that linguistic processes cannot be adapted to a deductive reasoning task instantaneously. Like most highly skilled processes, encoding, reencoding, and decoding processes are undoubtedly largely automatic; people do not have deliberate control over their internal workings. We assume that individuals can organize the flow of control through these processes to meet the demands of the task (e.g., first encode, then go back and reencode, and then decode) but that they have to take what they get whenever one is evoked. And although these outputs serve the needs of communication extremely well, they are less well suited to the demands of deductive reasoning. For example, in standard communication people often expect listeners to make plausible (if invalid) assumptions without being told explicitly. If someone asks you if you know what time it is, it goes without

saying that that person wants you to tell him or her (rather than just to answer yes or no). Conversely, listeners expect speakers to make important information explicit even if it is logically implied by what they have already said. The demands of deductive reasoning are quite different. Making unwarranted assumptions, however plausible, is precluded by correct deduction. Similarly, deduction requires appreciating the validity of certain information even if it was not made explicit. According to this view, the reason untrained participants make mistakes is that they are using linguistic processes that are adapted to the needs of communication rather than those of deduction. Although mistakes make these participants appear irrational, the errors arise from the decidedly rational strategy of applying the only available processes to a task for which they are poorly adapted (this argument is similar to a rational analysis as proposed by Anderson, 1990).

The verbal reasoning hypothesis does not imply that deduction is simply a special case of language processing. The demands of deductive tasks could lead participants to use their linguistic processes in very different ways than they are used in communication (e.g., repeatedly reencoding the problem statement and generating responses that relate specific terms from the problem). Although the verbal reasoning hypothesis proposes that the central processes in deduction are linguistic, these processes are assumed to be used in a way that reflects the demands of deduction rather than communication. Also, this hypothesis claims that the processes that are central to deduction are linguistic, not the representational structures. In particular, verbal reasoning does not preclude visuospatial representations or imagery (indeed, we follow Johnson-Laird, 1983, in assuming that people use a mental model representation). The claim is simply that the most important cognitive processes in deduction are the same ones that are used in language comprehension and generation. Furthermore, although the verbal reasoning hypothesis claims that high-level verbal processes (encoding, reencoding, and generation) play a central role in deduction, this does not imply that the detailed mechanisms underlying these processes are all linguistic. For example, these processes undoubtedly rely on underlying mechanisms (e.g., memory retrieval) that are used in a variety of nonlinguistic tasks. Rather, the hypothesis is that the same high-level processes, as a whole, that are used to transform verbal representations into semantic representations and back again also play a central role in deduction. Thus, *verbal* simply refers to transforming between verbal and semantic representations. Finally, the verbal reasoning hypothesis does not rule out reasoning-specific mechanisms, especially in participants who have had training in logic. People can certainly learn to use specific strategies (e.g., Venn diagrams and Euler circles) or even to change how they interpret specific words in deduction tasks (e.g., learning that "Some x" can refer to a set of xs that are distinct from any other xs in the problem; Lehman, Newell, Polk, & Lewis, 1993). But without such training, this hypothesis predicts that the major processes used in deduction will be linguistic rather than reasoning specific.

It is important to point out that the verbal reasoning hypothesis does not apply to tasks that require information beyond that presented verbally in the problem statement. The hypothesis is that on tasks in which all of the relevant information is provided verbally, the verbal processes of encoding and reencoding play a central role in reasoning itself (they are not simply transduction operators), at least for untrained participants. If this claim is true, then it is fair to say that deduction can usually be characterized as verbal reasoning, because almost all deductive tasks satisfy this criterion. There are exceptions, however. For example, tasks that require metadeduction, such as the Wason (1966) selection task, are not amenable to a pure verbal reasoning account. Linguistic processes may still play an important role in these tasks; however, because they require knowledge that cannot be extracted from the problem statement, linguistic processes will necessarily be insufficient. Similarly, a verbal reasoning explanation would presumably not apply to reasoning tasks that involve nonverbal materials. In particular, verbal reasoning does not predict that participants must translate nonverbal materials into a linguistic format to solve such problems.

In this article, we present evidence in favor of this view of deduction as verbal reasoning. We begin by showing that a computational model—VR—based on verbal reasoning can account for the detailed behavior of participants on a specific task, namely categorical syllogisms. VR handles all of the major syllogism variants, models all of the standard phenomena, and makes a number of novel predictions that have been empirically confirmed. Furthermore, it fits the behavior of individual participants with an accuracy that rivals the test–retest reliability of the participants themselves.[1] We then turn to previous explanations of the major phenomena of other deductive reasoning tasks and show either that they can be reinterpreted as verbal reasoning accounts or that verbal reasoning provides a more parsimonious explanation.

Verbal Reasoning on Categorical Syllogisms

Categorical syllogisms are reasoning problems consisting of two premises and a conclusion (Figure 1, left). There are four different premise types, called *moods,* each of which relates two of the terms and contains a quantifier (Figure 1, middle). The two premises always share a single term, called the *middle term* (*bowlers* on the left of the figure). A conclusion is legal if it relates the two noncommon terms (the *end terms: archers* and *chefs*) using one of the four moods. Sometimes there is no legal conclusion that is deductively valid. In that case, the correct response is that there is no valid conclusion (NVC). The first premise has two possible orders (relating x to y or relating y to x), as does the second premise (relating y to z or relating z to y); thus, together there are four possible orders, or *figures* (Figure 1, right). The four figures, together with the four moods for the first premise and the four moods for the second premise, combine to form 64 possible premise pairs. Different versions of the task require determining the validity of each member of a set of conclusions (Sells, 1936; Wilkins, 1928), choosing a valid conclusion from a set of alternatives (Ceraso & Provitera, 1971; Chapman & Chapman, 1959; Dickstein, 1975; Revlis,

[1] Thanks to Phil Johnson-Laird for graciously providing us with the raw data from his experiments with Bruno Bara and Mark Steedman that we used in these analyses.

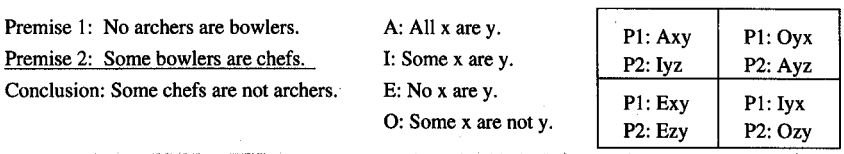

Figure 1. Categorical syllogism task (A = all, I = some, E = no, O = some not).

1975b), evaluating the validity of a given conclusion (Janis & Frick, 1943), and generating a valid conclusion given only the premises (Johnson-Laird & Bara, 1984).

Categorical syllogisms are relatively straightforward to understand and simple to administer. When responses are aggregated over groups of participants, some striking regularities emerge[2]:

1. The difficulty effect: The average participant makes many errors (often making errors on approximately half the tasks; Dickstein, 1975; Wilkins, 1928).

2. The validity effect: The average participant performs better than chance (Johnson-Laird & Steedman, 1978; Revlis, 1975b).

3. The atmosphere effect: Excluding NVC responses, (a) if either premise is negative (*No x are y* or *Some x are not y*), most responses are negative, and otherwise most are positive, and (b) if either premise is particular (referring to a subset: *Some x are y* or *Some x are not y*), most responses are particular, and otherwise most are universal (*All x are z* or *No x are z*; Sells, 1936; Woodworth & Sells, 1935).

4. The conversion effect: Excluding NVC responses, many erroneous responses would be correct if the converse of one or both premises were assumed to be true (Chapman & Chapman, 1959; Revlis, 1975b).

5. The figural effect: Excluding NVC responses, if only one end term (x or z) appears as the subject of a premise, that term tends to appear as the subject of the conclusion (Johnson-Laird & Bara, 1984; Johnson-Laird & Steedman, 1978).

6. The belief bias effect: Participants are more likely to generate and accept as valid a conclusion that they believe to be true in comparison with one they believe to be false, independent of its true logical status (Evans, Barston, & Pollard, 1983; Janis & Frick, 1943; Morgan & Morton, 1944; Oakhill & Johnson-Laird, 1985; Oakhill, Johnson-Laird, & Garnham, 1989; Wilkins, 1928).

7. The elaboration effect: Participants are more accurate if the premises are elaborated to be unambiguous (e.g., *All A are B, but Some B are not A;* Ceraso & Provitera, 1971).

VR: A Computational Model for Categorical Syllogisms

VR is a computational model of behavior on categorical syllogisms that is based on the idea of verbal reasoning. Figure 2 presents the general control structure of VR when generating a conclusion from a pair of premises, as well as the major assumptions of the model.[3]

In keeping with verbal reasoning, the central processes in VR are linguistic: initial encoding, conclusion generation, and reencoding. In contrast, VR's control structure reflects the basic demands of the reasoning task. At the very least, the task requires encoding the premises (initial encoding) and producing a response (generating a conclusion). If VR fails in generating a conclusion, then it repeatedly reencodes the premises until generation succeeds or until it gives up. Reencoding in this context is a natural response to the task demand of producing a conclusion: The system lacks the knowledge to relate the end terms, the main source of knowledge for the problem is the premises, and encoding is the most natural way of extracting knowledge from a premise. If repeated reencoding fails to lead to a conclusion, then VR gives up and responds NVC. We now turn to specifying each of the processes in Figure 2: initial encoding, conclusion generation, reencoding, and giving up.

Initial encoding. Following Johnson-Laird (Johnson-Laird, 1983; Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991), VR first constructs a mental model of a situation in which the premises are true (as previously discussed, VR assumes the same data structure as mental model theory, but the central processes proposed by the theories—reencoding vs. searching for alternative models—are different). Mental models consist of a set of model objects with properties and relations among them. Objects, properties, and relations in a mental model are representational constructs that correspond to objects, properties, and relations in the situation being represented (the referent). Indeed, the defining characteristic of mental models is that they satisfy the structure-correspondence principle: All objects, properties, and relations in the mental model must map one to one into objects, properties, and relations in the referent. VR uses an annotated model representation in which properties can be annotated with two additional pieces of information: a *not* flag (represented by the − symbol) indicating that the object does not have the specified property and (b) an *identifying* flag (represented by the prime [′] symbol) indicating that the object is identified by the specified property (e.g., VR distinguishes a painter [identifying] who is a sculptor from a sculptor [identifying] who is a painter). Identifying properties are distinguished from other, secondary properties by being more easily accessible. Specifically, when VR generates conclusions based on its annotated models, it produces conclu-

[2] Wilkins (1928) and Sells (1936) have also claimed evidence for a concreteness effect: Participants are more accurate on syllogisms involving concrete (*archer, bowler*) rather than abstract (*A, B, C*) terms. The effect is not well established, however (attempted replications have failed; Gonzalez-Marques, 1985), and it may disappear when believability effects are controlled (Revlis, 1975a; Revlis, Ammerman, Petersen, & Leirer, 1978).

[3] Variants of this model have been built that handle all of the standard task variants (e.g., testing conclusions and multiple choice). For more information, contact Thad A. Polk.

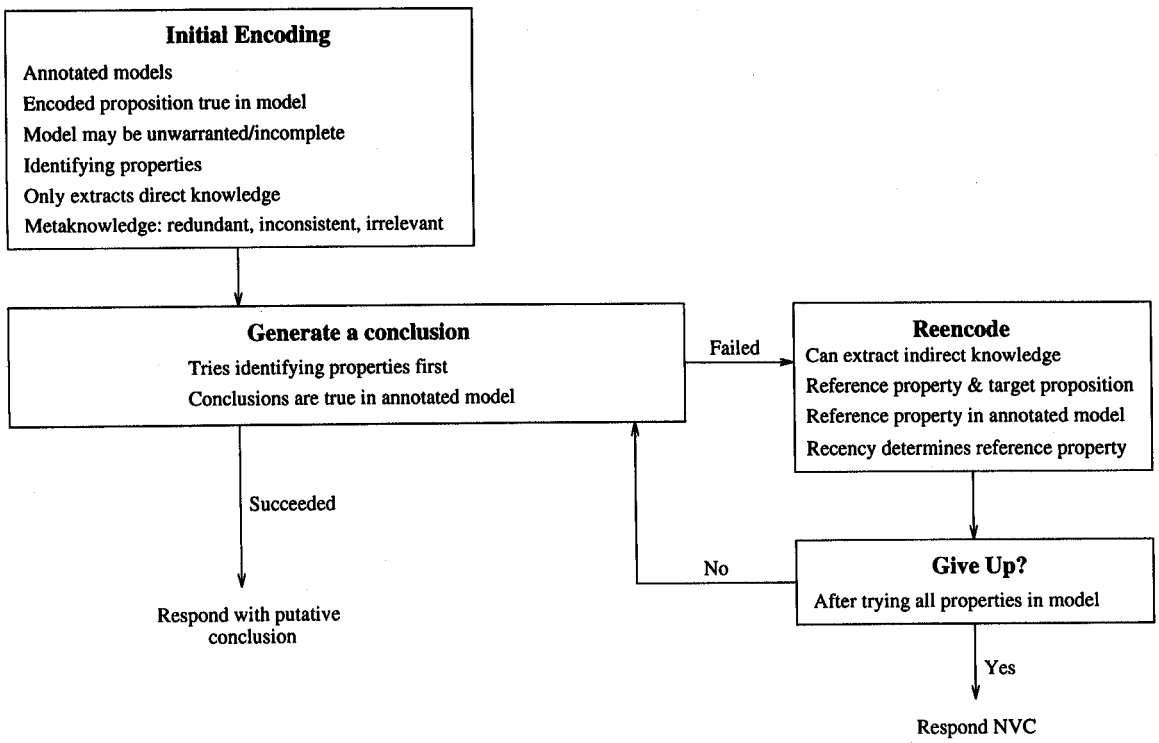

Figure 2. Summary of VR (NVC = no valid conclusion).

sions about identifying properties before trying conclusions about secondary properties. For syllogisms, the identifying properties simply correspond to the topics of the propositions being encoded (i.e., their grammatical subjects).

After encoding a proposition, the resulting annotated model is guaranteed to represent a situation in which that proposition is true. But, in general, there are an infinite variety of such situations, and most of them are unwarranted or incomplete with respect to the initial proposition. That is, an annotated model may encode information that is not inherent in a proposition (and be unwarranted) or may fail to encode information that is inherent (and be incomplete). Figure 3 gives examples of each. On the left is an annotated model that is unwarranted with respect to the proposition *Some archers are bowlers.* The annotated model consists of two model objects represented as a set of properties enclosed in parentheses. The first object corresponds to a person who is an archer and bowler, but the second is an archer and explicitly not a bowler. The proposition *Some archers are bowlers* is true in this annotated model, but the model also encodes information that is unwarranted given only that proposition: the fact that some archers are not bowlers. On the right is an example of an incomplete annotated model. Once again, the model supports the initial proposition (*No archers are bowlers* in this case), but now it fails to reflect the fact that the second object (the bowler) cannot be an archer. In this case, then, the annotated model is incomplete with respect to the initial proposition.

During initial encoding, VR encodes only direct knowledge, that is, information about the topic of each proposition. For example, suppose encoding the first premise of a syllogism leads to an annotated model with a single object (A' B). Then an initial encoding of *No C are B* will not affect that object because the direct knowledge is about C and that model object does not have property C. Only if this direct knowledge is insufficient to produce a conclusion will VR try to extract indirect knowledge (i.e., knowledge about nontopic properties) via reencoding.

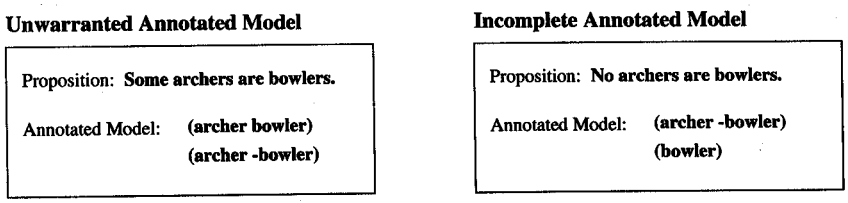

Figure 3. Examples of unwarranted and incomplete annotated models.

	No object (X ...)	An object (X ...) exists
All x are y	(X' Y)	all (X ...) → (X' ... Y)
Some x are y	(X' Y) (X')	MR (X ...) → (X' ... Y) (X' ...)
No x are y	(X' -Y)	all (X ...) → (X' ... -Y)
Some x are not y	(X' -Y) (X')	MR (X ...) → (X' ... -Y) (X' ...)

Figure 4. Default encodings for VR. At left are the model objects that would be added to the annotated model if no existing model objects relate to the topic of the proposition (X). If the annotated model already contains model objects with property X, shown at right are how those objects would be changed (e.g., augmenting all such objects with property Y or augmenting the most recently accessed such object with property Y). ' = identifying property; MR = most recently accessed; . . . = other properties.

Finally, when encoding a proposition into an annotated model, VR encodes metaknowledge representing whether the proposition was redundant, inconsistent, or irrelevant with respect to the annotated model. The choice of flag is based on how the annotated model was affected during encoding: If it was unchanged, then the proposition was redundant; if an existing property was removed or replaced, then the proposition was inconsistent; and if the existing objects were unchanged but new unrelated objects were created that share no properties with the old objects, then the proposition was irrelevant.

Figure 4 shows default encodings consistent with these assumptions for each of the four syllogism premise moods. For each premise type, the figure shows how the model would be augmented, both when there are no objects in the model that relate to the topic of the proposition (left) and when there are (right). We assume that there is substantial individual variation in these encodings, and later we address this diversity.

Conclusion generation. Once VR has encoded the premises, it attempts to produce a conclusion based on its annotated model. It does so using a simple generate-and-test procedure. First, it proposes simple propositions that it considers to be true based on the annotated model, and then it tests each proposed proposition to determine whether it is a legal syllogism conclusion (i.e., whether it relates the end terms using one of the four standard moods [*All, Some, No,* or *Some not*]). The simple propositions that VR may propose include the legal syllogism conclusions but may also involve the middle term, other quantifiers, or both. As previously mentioned, VR proposes propositions about identifying properties first. The proposal phase is based on templates that are associated with each simple proposition. Each such template specifies the conditions under which the associated proposition should be considered true (and hence proposed) in the annotated model. When the annotated model matches the template, the associated simple proposition is proposed. If none of the proposed propositions are legal syllogism conclusions, generation fails and reencoding is evoked. If at least one is legal, generation succeeds. In practice, there is rarely more than one proposed conclusion that is legal. When there is, VR produces the set of conclusions it considers equally probable. It could obviously choose randomly among the alternatives, but having the entire set makes it possible to compute the exact expected value of accurate predictions.

As a default, we assume that *All x are y/No x are y* will be proposed when there is an object with X as an identifying property and all objects with property X also have property $Y/-Y$. *Some x are y/Some x are not y* will be proposed when there is an object with X as an identifying property and there is at least one object with properties X and $Y/-Y$. Again, we address individual differences later.

Reencoding. If generation fails to produce a legal conclusion, VR tries to extract additional knowledge from the premises by reencoding. Unlike initial encoding, reencoding can extract indirect knowledge about nontopic properties. It first chooses a property based on the annotated model (the reference property) and then tries to extract additional knowledge about that property from any premise that mentions it (the target proposition). If the reference property is the topic of the target proposition, then reencoding works just like initial encoding (extracting only direct knowledge). If, however, the reference property is not the target proposition's topic, then VR will try to extract indirect knowledge about that property from the proposition. For example, in trying to extract more knowledge about property B, VR might reencode the premise *No A are B* (assuming that generation has failed so far). This premise does contain valid knowledge about B (that objects with property B in the model do not have property A), and VR may or may not be able to extract it. VR selects reference properties on the basis of how recently the corresponding model objects were accessed by the system. It starts with the most recently accessed objects because they are less likely to reflect information from older propositions. Reencoding with respect to these properties is more likely to augment the annotated model and lead to progress.

Giving up. We assume that most participants give up only as a last resort, when they feel confident that repeated attempts to extract more knowledge will not lead anywhere. VR reaches such a point after it has tried extracting more knowledge about every property in the annotated model. This is when it quits and responds NVC.

An example of VR's behavior. Figure 5 illustrates VR's default behavior on the following syllogism: *Some B are A. All B are C. Therefore, what necessarily follows?* As a means of illustrating reencoding, this version of VR has been given the ability to extract some indirect knowledge: When reencoding *Some x are y* with respect to property y, it augments the model in the same way that it would when encoding *Some y are x.*

VR's behavior on this syllogism follows the structure presented in Figure 2. After initially encoding the premises (Step 1 in Figure 5), it repeatedly reencodes them with respect to different reference properties until it is able to generate a legal conclusion *Some A are C* as its response (Steps 2–4).

Consider the initial encoding in more detail (Step 1). VR begins by encoding the first premise *Some B are A* into an annotated model. In keeping with the default encoding for *Some x are y,* VR creates two new objects: (B') and (B' A). (Throughout this example, the model objects in Figure 5 are

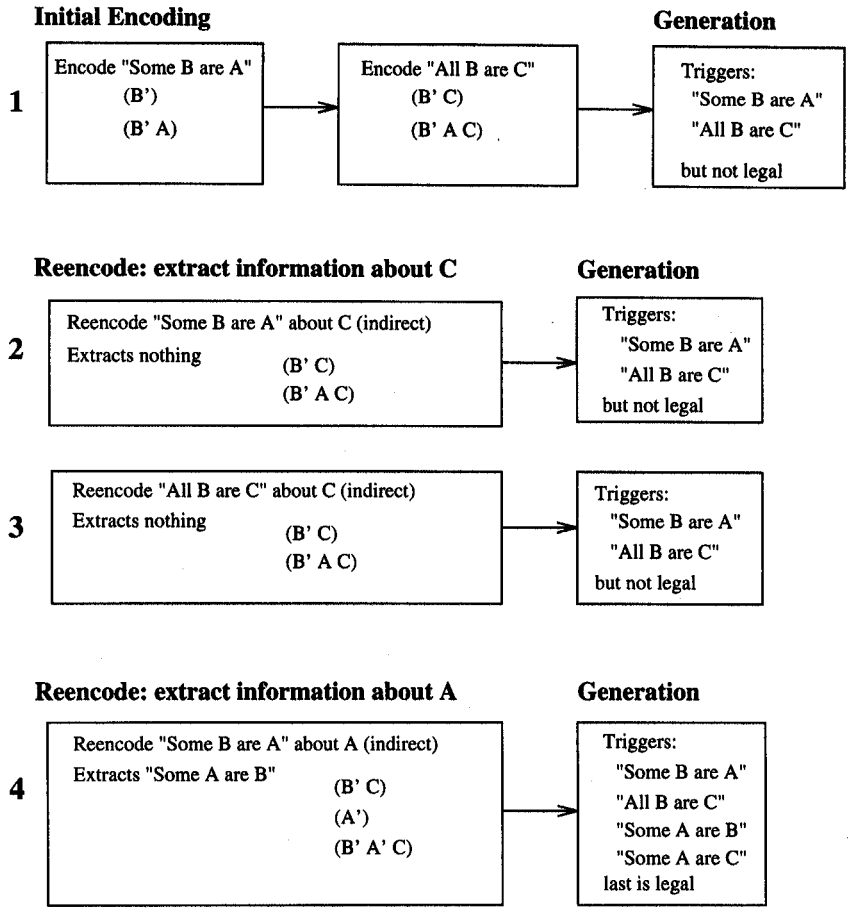

Figure 5. VR on *Some B are A, All B are C*. ' = identifying property.

ordered from the least recently accessed [at the top] to the most recently accessed [at the bottom].) VR then encodes the second premise *All B are C*. It marks property *B* as identifying in all objects (a null operation in this case) and augments all objects having property *B* with property *C*. In this example, both objects are augmented. The result is an annotated model with two model objects (*B' C*) and (*B' A C*). This annotated model satisfies the default generation templates for *Some B are A* and *All B are C*, and so these conclusions are proposed. Because neither is a legal syllogism conclusion (they do not relate the end terms), VR must resort to reencoding. The most recently accessed model object is the recently augmented (*B' A C*), and *C* is its most recent property, so VR will begin reencoding using *C* as the reference property. Property *A* is the next most recent, and so it will be tried next, followed by *B*.

Steps 2 and 3 illustrate VR's reencoding when the reference property is *C*. Reencoding *Some B are A* extracts nothing about *C* (Step 2). There is no way it could, because the premise does not even mention that property. The annotated model therefore remains unchanged, and the same (illegal) conclusions are proposed. *All B are C* does mention property *C*, and so there is a possibility that reencoding will extract new indirect information from this premise. In this example, however, VR does not realize it can extract anything about *C* from the premise, and so once again the model is unchanged and generation fails (Step 3).

VR then turns to reencoding the premises using reference property *A* (Step 4). Unlike the previous attempts, reencoding *Some B are A* does extract some indirect knowledge, as previously mentioned. In this context, VR extracts the knowledge *Some A are B*. More precisely, it augments the annotated model in the same way it would when encoding that proposition. It marks property *A* as identifying and creates a new object (*A'*). Because *A* has now become an identifying property, generation finally proposes conclusions about *A* in addition to those about *B*. One of these new conclusions (*Some A are C*) is legal, and so generation succeeds and reencoding stops.

In keeping with verbal reasoning, the central processes in VR are linguistic: encoding, reencoding, and generation. VR does not have any processes devoted specifically to deduction (e.g.,

Table 1
Percentages of Correct, Atmospheric, Conversion, and Figural Responses From Humans, VR, and Random Data

Response type	Human data							VR's predictions				
	Unlimited (n = 20)	Timed (n = 20)	Revised (n = 20)	Week 1 (n = 20)	Week 2 (n = 20)	Inder (n = 3)	Total (n = 103)	VR1–3	VR1	VR2	VR3	Random
Correct	40	49	45	59	69	61	53	58	64	75	38	15
Atmospheric	69	72	76	81	84	84	77	89	100	100	83	25
Conversion	28	38	38	38	51	54	37	40	50	29	41	9
Figural	90	81	83	90	86	70	86	89	100	86	86	50

Note. VR = verbal reasoning model.

applying inference rules or searching for alternative models). Can such a model accurately capture the behavior of real participants? We now analyze VR's ability to fit aggregate and individual data as well as to make accurate novel predictions.

Aggregate Data

Analysis 1: Phenomena using standard content and form. Table 1 presents aggregate data from six experiments and results using three default versions of VR as well as their average. Standard syllogism variants were used in which effects of content (belief bias) and form (elaboration effect) did not play a role. At the far right are the values that would be expected in purely random data. The top row in the table presents the percentage of legal responses that were correct (of 6,592 total participant responses, 159 [2.4%] were either illegible or did not correspond to one of the nine legal responses [many involved the middle term], and these were not included). The other three rows present measures of the atmosphere, conversion, and figural effects. Because the atmosphere effect does not apply to NVC responses, these were not included. Instead, the numbers represent the percentage of legal responses other than NVC that followed the atmosphere effect. The conversion effect also does not apply to NVC responses, and, in addition, it is restricted to errors. Consequently, the numbers in the third row represent the legal but erroneous non-NVC responses that would be valid if the converse of one or both premises were true. Finally, the figural effect is relevant only to 32 of the 64 tasks (those in the *xy–yz* and *yx–zy* figures), and it too does not apply to NVC responses. The numbers in the bottom row therefore represent the percentage of legal non-NVC responses on relevant tasks that followed the figural effect.

The first three data sets were collected from students at the University of Milan (Johnson-Laird & Bara, 1984). In the first (unlimited), participants were given an unlimited amount of time to write down a response. In the second (timed), participants were asked to respond within 10 s. Afterward, they were given the same tasks along with the responses they provided in the timed condition and were given 1 min to revise their answers if desired. These data are recorded in the fourth column (revised). The next two data sets (Week 1 and Week 2) were collected a week apart from the same group of students at Teachers College, Columbia University. The last data set (Inder) was collected from 3 undergraduates at the University of Edinburgh (Inder, 1986, 1987). None of the participants had training in logic.

The three columns labeled VR1, VR2, and VR3 present predictions from three versions of VR, and column VR1–3 presents their average. These three systems were intended to be very direct implementations of the defaults described previously. VR1 is the simplest possible such system, and so it is very streamlined. In particular, it does not extract any indirect knowledge when reencoding and often fails to relate the end terms, producing a large number of NVC responses (52 of 64). As a result, the measures of the atmosphere, conversion, and figural effects for VR1 may not be reliable because they are based on a small sample of non-NVC responses (12 of 64 [19%]). To overcome this problem, we built two other versions of VR that do extract indirect knowledge and that, consequently, produce fewer NVC responses. VR2 is identical to VR1 except that it can extract two pieces of (correct) indirect knowledge: (a) When the target proposition is *Some x are y* and the reference property is *y*, it augments the model in the same way that encoding the proposition *Some y are x* would augment it, and (b) when the target proposition is *No x are y* and the reference property is *y*, it augments the model in the same way that *No y are x* would.[4] This version of VR produces significantly more non-NVC responses (25 of 64) than VR1. VR3 is an extreme case that always extracts indirect knowledge and therefore produces relatively few NVC responses (8 of 64). Appendix A presents the indirect knowledge that VR3 extracts from all combinations of target propositions and reference properties. The specific indirect knowledge extracted was chosen to be consistent with the target proposition, to relate the same two terms, and to share the same quantity (universal or particular). Appendix B presents the specific responses predicted by each of the default versions of VR.

The numbers in the far right column of Table 1 (random) correspond to percentages that would be expected by chance alone. There are 576 legal task–response pairs (64 tasks × 9 legal responses each). Of these, 85 (15%) represent a correct response to the associated task. This number is higher than the

[4] Encoding indirect knowledge could work either by generating an intermediate proposition (e.g., *Some y are x* and *No y are x* in these cases) and then encoding that proposition (an inner speech strategy) or by encoding the indirect knowledge directly without an intermediate proposition. VR is consistent with either mechanism.

total number of tasks (64) because some tasks have more than a single correct response. If all 64 NVC responses are removed, 512 task–response pairs remain that do not involve NVC. Of these, 128 (25%) are consistent with the atmosphere effect. Similarly, there are 464 task–response pairs that both represent errors and do not involve NVC. Forty of these (9%) would be valid if the converse of one or both premises were true. Finally, the figural effect does not apply to 256 of the 512 non-NVC task–response pairs (those in the *yx–yz* and *xy–zy* figures). Of the remaining 256, 128 (50%) follow the figural effect.

The human data in Table 1 reflect the first five of the seven regularities listed earlier. The difficulty of the task is reflected in the low percentage of correct responses (53% on average). Nevertheless, participants performed far better than would be expected by chance alone (15% in the column on the far right), demonstrating a validity effect. In keeping with the atmosphere effect, 77% of the legal non-NVC responses were consistent with atmosphere (25% would be expected by chance). A conversion effect is also apparent because 37% of the legal, erroneous non-NVC responses would have been valid with conversion, in comparison with only 9% in random data. Finally, the data also reflect a strong figural effect; the percentage of relevant, legal non-NVC responses that followed the figure of the premises in the human data (86%) was much higher than what would be expected in random data (50%). Note that these five regularities are reflected in all six data sets, indicating their robustness.

Table 1 also demonstrates that all three default versions of VR produce these five regularities. They solve only between 38% and 75% of the syllogisms correctly (difficulty effect), but this is much better than would be expected at random (validity effect). Also, far more of their responses follow the atmosphere (83%–100% of non-NVC responses), conversion (29%–50% of non-NVC errors), and figural effects (86%–100% of relevant non-NVC responses) than would be expected in random data (25%, 9%, and 50%, respectively). The crucial point is that computational models based on the verbal reasoning hypothesis do account for these empirical phenomena.

These versions of VR not only produce the regularities, but the qualitative size of the effects is similar to that in the human data. The mean percentages correct in the human data sets from Table 1 range from 40% to 69%, with a mean of 53%; for the three VR systems, they range from 38% to 75%, with a mean of 58%. Similarly, the size of the atmosphere effect in these data sets (ranges from 69% to 84%, with a mean of 77%) is similar to that produced by the VR systems (83% to 100%, with a mean of 89%), as is the size of the conversion effect (28% to 54% [M = 37%] in human data and 29% to 50% [M = 40%] for the VR systems) and the figural effect (70% to 90% [M = 86%] in human data and 86% to 100% [M = 89%] for the VR systems).

Because VR predicts these regularities, analyzing how it does so will suggest explanations for why they occur. First consider the question of why syllogisms are so difficult. VR suggests that the main reason is that language comprehension is not guaranteed to deliver a necessary and sufficient representation of what the premises are about. VR simply constructs a model of a situation that is consistent with the premises (i.e., the premises will be true in the model). Such a model can both support conclusions that are not strictly valid (and be unwarranted) and fail to support conclusions that are (and be incomplete). Figure 3 gave examples of each.

The assumption that annotated models are always consistent, if not always necessary and sufficient, also provides an explanation of the validity effect (the fact that people perform better than chance). If an annotated model is consistent with the premises, then it follows that no entailments of the premises can be false in the model (this does not mean that the entailments must be true in the model). Otherwise one (or both) of the premises would have to be false. Equivalently, a consistent annotated model never supports conclusions that contradict an entailment of the premises. So if people base their conclusions on an annotated model that is consistent with the premises, their responses will always be possible (although not necessary); people will not consider (invalid) conclusions that contradict an entailment of the premises. Consequently, they do better than just randomly selecting from all possible conclusions.

The reason that VR produces the atmosphere effect and, hence, the suggested explanation is that the standard semantics of the premises rule out most nonatmospheric conclusions. Assuming that *some* is typically interpreted as *some but not necessarily all* (or just *some but not all*) explains why particular responses are so common for syllogisms that involve a particular premise (but not for those that do not). When a *some* premise is read, it creates additional model objects that do not satisfy the predicate of the premise. These additional model objects rule out universal conclusions and lead to particular responses. For example, the default encodings lead to the following annotated model for the premise *Some A are B:*

(A')
$(A'\ B)$

where the upper model object encodes the *not all* information. Augmenting this model based on *All B are C* leads to

(A')
$(A'\ B'\ C)$

and the upper model object refutes the universal conclusion *All A are C,* leading to a *some* response instead. If both premises are universal, these additional objects are not created, and universal conclusions can be drawn. If the interpretation of *some* during generation is similar to that during comprehension (*some but not* [*necessarily*] *all*), then particular conclusions will be drawn only if universals cannot. Consequently, two universal premises tend to produce a universal conclusion.

In a similar way, negative premises tend to refute positive conclusions because they create negative rather than positive associations between properties. For example, consider reading *All A are B* and *No B are C* using the default encodings. *All A are B* leads to the single model object ($A'\ B$), and then *No B are C* augments it with $-C$—($A'\ B'\ -C$)—creating a negative rather than positive association. Because negative premises tend to produce negative associations, the resulting conclusions are also negative. Conversely, if both premises are positive, then positive rather than negative associations among properties are created, and conclusions tend to be positive.

There are two main reasons VR produces a conversion effect.

The first is straightforward: VR often extracts indirect knowledge that is equivalent to the converse of a premise (even if that converse is invalid). For example, when VR3 tries to extract knowledge about y from a premise that relates x to y, the knowledge it encodes is always equivalent to the converse of the premise (see Appendix A). In two of the four cases, the converse is invalid and can lead to errors consistent with the conversion effect. Consider the processing that VR3 goes through while working on the following premise pair: *Some A are B, All C are B*. Initial encoding leads to the following annotated model:

(A')
$(A'\ B)$
$(C'\ B)$

and B is chosen as the first reference property. Reencoding the first premise with respect to B changes only the recency of some of the objects, but reencoding *All C are B* with respect to B augments the model in the same way encoding *All B are C* would:

(A')
$(C'\ B')$
$(A'\ B'\ C)$

which supports both *Some A are C* and *Some C are A*. Both responses are consistent with the conversion effect (i.e., they are invalid but would be correct if the converse of one or both premises were true).

This explanation is reminiscent of previous theories of syllogistic reasoning that assumed illicit conversion (Chapman & Chapman, 1959; Revlis, 1975b), but there are important differences. For one, those theories attempted to explain most non-NVC errors in terms of illicit conversion, but we assume that there are a variety of other sources of error. Indeed, as shown in Table 1, only 40% of non-NVC errors produced by VR1, VR2, and VR3 can be explained in terms of illicit conversion, but this simulates the human data very well (in which only 37% of non-NVC errors were consistent with the conversion effect). The fact that approximately 60% of non-NVC errors cannot be explained in terms of illicit conversion poses problems for the Chapman and Chapman (1959) and Revlis (1975b) theories but not for VR. Furthermore, Revlis intentionally implemented the strongest version of the conversion hypothesis, namely, that both premises are always explicitly converted and that both they and their converses are encoded at the same time (as Revlis pointed out, Chapman and Chapman were vague on this point). VR makes no such assumptions. In VR, indirect knowledge is extracted only if direct knowledge fails to lead to a conclusion. And although indirect knowledge is assumed to augment the annotated model in the same way encoding a corresponding explicit proposition would, VR does not assume that the indirect knowledge is available as an explicit proposition. It may or may not be; VR does not take a stand on the issue.

VR1 and VR2 both produce conversion effects (Table 1), and yet neither extracts invalid indirect knowledge (VR2 extracts only valid indirect knowledge, and VR1 does not extract any at all). According to VR, then, there must be at least one other factor that contributes to errors consistent with the conversion effect. The source of most of these other errors is the traditional fallacy of the undistributed middle term, which shows up in the interpretation of particular premises. Consider the premise pair *All A are B, Some B are C*. Both VR1 and VR2 behave the same way on this problem. Encoding the first premise creates a single model object $(A'\ B)$, and the second premise augments it and creates a new one as well:

$(A'\ B)$
$(A'\ B'\ C)$

This annotated model leads to the invalid conclusion *Some A are C*. Furthermore, this conclusion is consistent with the conversion effect because it would be valid if the converse of *All A are B* (i.e., *All B are A*) were true. But no knowledge corresponding to the converse of either premise has been extracted (indeed, no indirect knowledge has been extracted at all). The problem is that *Some B* was interpreted as referring to a subset of the objects mentioned in the first premise, and this interpretation is invalid (to see this, consider the syllogism *All cats are animals, Some animals are dogs. Therefore, some cats are dogs*). So some errors consistent with the conversion effect can be explained in terms of the erroneous interpretation of particular premises, without any reference to illicit conversion.

VR's explanation of the figural effect is based on the special availability of proposition topics in the annotated model. After encoding, properties in the annotated model that correspond to premise topics are more available than other properties (they are marked as identifying properties). During generation, conclusions about those properties are tried first. If only one end term appeared as the topic of a premise, then the corresponding property will be marked as identifying, and generation will tend to produce conclusions about it. In the *xy–yz* and *yx–zy* figures, this leads to *xz* and *zx* conclusions, respectively, as predicted by the figural effect.

Analysis 2: Effects of belief. Tables 2 and 3 present percentages of trials on which the provided or suggested conclusion was considered valid, both when that conclusion was consistent with beliefs [column B (%) in the tables] and when it was inconsistent with beliefs [column D (%) in the tables]. In most cases, believability ratings were collected from an independent set of participants. Table 2 presents four experimental paradigms involving testing conclusions, and the four in Table 3 involve generating conclusions.[5]

The second and third columns in Table 2 present the human data from four belief bias experiments in which provided conclusions had to be evaluated. The other columns present the predictions of three versions of VR, as well as their average. These systems are the same as those described previously but have been modified to respond to the new experimental paradigm. In particular, they test rather than generate conclusions. They do so by encoding the premises followed by the conclusion and basing their responses on metaknowledge. If the conclusion is found to be inconsistent with the annotated model of the

[5] Wilkins (1928), Janis and Frick (1943), and Morgan and Morton (1944) also collected data relevant to belief bias, but these studies did not include the critical comparisons between believed and disbelieved conclusions for syllogisms with the same logical form.

Table 2
Percentages of Believed and Disbelieved Conclusions Accepted as Valid

	Humans		VR1–3 average		VR1		VR2		VR3	
Experiment	B(%)	D(%)	B(%)	D(%)	B(%)	D(%)	B(%)	D(%)	B(%)	D(%)
Evans et al. (1983)										
Experiment 1	92	50	67	33	50	0	75	50	75	50
Experiment 2	76	38	63	25	50	0	75	50	63	25
Experiment 3	79	30	65	29	50	0	75	50	69	38
Oakhill et al. (1989)										
Experiment 3	58	52	75	50	75	50	75	50	75	50

Note. B(%) indicates the percentage of believed conclusions that were accepted as valid; D(%) indicates the same percentage for disbelieved conclusions. VR = verbal reasoning model.

premises, they consider it invalid. If it is redundant, they consider it valid. If neither of these outcomes occurs, then they repeatedly reencode until they can decide or until they give up. We simulated the belief effects by having VR choose randomly between "invalid" and the response suggested by belief whenever it gave up (i.e., encoding and reencoding were inconclusive). We did not incorporate any belief effects when the results of reasoning were conclusive.

The four experiments presented in Table 3 investigated belief bias effects involved in generating a conclusion rather than determining the validity of one that was given. In these studies, participants were presented with premise pairs and the experimenters manipulated the believability of the conclusion "suggested" by the premises. The suggested conclusions were selected on the basis of Johnson-Laird's theories of syllogistic reasoning (Johnson-Laird, 1983; Johnson-Laird & Bara, 1984)—they corresponded to conclusions that are true in one mental model of the premises (the specific mental model that Johnson-Laird & Bara, 1984, assumed to be built first)—and were confirmed to be common responses in the data he collected. Table 3 shows the percentage of generated conclusions that matched the suggested conclusion when that conclusion was both believable (46%) and unbelievable (30%).

The second and third columns in Table 3 present the human data from these four experiments. The other columns present the predictions from the three versions of VR and their average. Once again, the systems are the same as before but have been adapted to the demands of the task; they generate conclusions just like the original versions, but they incorporate a belief bias. Specifically, whenever the systems generate a conclusion that is unbelievable, they go back and try to generate a different conclusion by repeatedly reencoding the premises. If that fails, they choose randomly between their original conclusion and NVC.

Tables 2 and 3 demonstrate a belief bias in both testing and generating conclusions. In all four experiments in Table 2, the percentage of believed conclusions that were accepted as valid was consistently larger than the percentage for disbelieved conclusions (92% vs. 50%, 76% vs. 38%, 79% vs. 30%, and 58% vs. 52%). This same effect can be seen in the four experiments that required generating rather than testing conclusions (Table 3).

Furthermore, Tables 2 and 3 show that the three VR systems are more likely as well to accept and generate a conclusion if it is believed rather than is disbelieved, demonstrating that the model also produces a belief bias. In the one experiment in which the systems failed to show an effect of beliefs (Oakhill & Johnson-Laird, 1985, Experiment 1), only two premise pairs

Table 3
Percentages of Generated Conclusions That Matched Believed and Disbelieved Suggested Conclusions

	Humans		VR1–3 average		VR1		VR2		VR3	
Experiment	B(%)	D(%)	B(%)	D(%)	B(%)	D(%)	B(%)	D(%)	B(%)	D(%)
Oakhill & Johnson-Laird (1985)										
Experiment 1	46	30	0	0	0	0	0	0	0	0
Experiment 2	49	43	50	25	0	0	100	50	50	25
Oakhill et al. (1989)										
Experiment 1	66	37	50	17	50	25	50	17	50	8
Experiment 2	74	28	50	15	50	25	50	13	50	6

Note. B(%) indicates the percentage of generated conclusions that matched a believed suggested conclusion; D(%) indicates the same percentage when the suggested conclusion was disbelieved. The suggested conclusion is what would be predicted by Johnson-Laird and Bara (1984) if the participant considered only one mental model. VR = verbal reasoning model.

Table 4
Percentage Correct for Elaborated and Unelaborated Premises in Humans and VR

Premise type	Humans	VR1–3 average	VR1	VR2	VR3
Unelaborated	58	59	62	69	46
Elaborated	80	95	100	100	85

Note. VR = verbal reasoning model.

were used. In both cases, VR1 and VR2 produced NVC and VR3 produced *Some C are not A.* These responses did not correspond to the suggested conclusion *Some A are not C,* and, as a result, VR did not consider them either believable or unbelievable. Consequently, there was no opportunity for an effect of belief to show up. Similarly, the second experiment involved only two syllogisms, and VR1 produced NVC on both; thus, it did not produce a belief bias. In all other cases (and in all cases involving testing given conclusions), however, all three systems did produce a belief bias effect.

At the most general level, there are two reasons VR is biased by beliefs. First, its reasoning attempts are often inconclusive, so it looks for other knowledge sources on which to base a decision. Beliefs are such a knowledge source. More specifically, beliefs can influence how VR responds after giving up (by responding on the basis of beliefs rather than in keeping with a default). Second, because VR's reasoning attempts can be faulty, it will reconsider results it would normally accept as valid if they contradict beliefs. For example, beliefs can influence the criteria under which conclusion generation is considered successful (the "succeeded" test for conclusion generation in Figure 2). If VR generates a conclusion that contradicts beliefs, then it will go back and try to generate a different one, even if the original conclusion was legal.

It is important to point out that these explanations depend on assumptions beyond the verbal reasoning hypothesis. The reason is that the effect depends on information that is not provided in the problem statement (participants' beliefs). Consequently, verbal processes will necessarily be insufficient to produce these effects. Of course, verbal reasoning is still apparent in the behavior of these systems (e.g., they repeatedly reencode propositions both to test conclusions and to generate new ones). But verbal reasoning provides only part of an explanation. Additional assumptions are necessary for a complete explanation.

Analysis 3: Elaboration effect. Table 4 presents data relevant to the elaboration effect. The top row presents the percentage of responses that were correct for a set of 13 standard syllogisms used by Ceraso and Provitera (1971). The bottom row presents the percentage of correct responses for the same syllogisms when their premises were elaborated to be unambiguous. The elaborated and unelaborated forms of the premises are shown in Figure 6. In the unelaborated condition, only the premises in boldface were presented. In the elaborated condition, the others were included as well. Each syllogism was presented with four alternatives: (a) *All A are C,* (b) *Some A are C,* (c) *No A are C,* and (d) *Can't say.* The participants were asked to choose the valid response from among these four.

The human data (second column in Table 4) were collected from 80 students at the Newark campus of Rutgers University (Ceraso & Provitera, 1971). The other columns present predictions from the three default versions of VR and their average. For this analysis, the VR systems were unchanged. The only difference between the conditions was whether the systems were given only the unelaborated premises or the entire set of elaborated premises (Figure 6).

Table 4 illustrates the elaboration effect. The percentage of responses that were correct rose significantly when the premises were elaborated to be unambiguous (58% vs. 80%). Table 4 demonstrates that the VR systems exhibited an elaboration effect as well. All three systems produced a higher percentage of correct responses when the premises were elaborated to be unambiguous (between 85% and 100%) than when they were left unelaborated (46% to 69%). Furthermore, the percentage of elaborated and unelaborated syllogisms that the three VR systems solved correctly was also quite similar to the human data (59% vs. 58% for unelaborated premises and 95% vs. 80% for elaborated premises). The bottom line is that systems based on the verbal reasoning hypothesis do account for the elaboration effect.

VR's behavior on the premises in Figure 6 suggests that part of the elaboration effect on these tasks is due to artifacts in the experiment. Specifically, there are two main problems with the Ceraso and Provitera (1971) study that could account for part of the elaboration effect they observed: It presented only four responses from which to choose, and the valid conclusions were sometimes different for the elaborated and unelaborated premises. Consider Tasks 8 and 12 in Figure 6. The VR systems produce the erroneous response *Some A are C* to the unelaborated forms of both tasks but give the correct response (NVC) in the elaborated forms. These systems do generate one conclusion (*Some C are not A*) while working on the elaborated premises; however, because this conclusion is not one of the alternatives and because they cannot produce any others, they respond with NVC. In a different version of the task (that included this conclusion as a legal choice), these systems would also have solved the elaborated tasks incorrectly. Task 4 (Figure 6) demonstrates a different problem with the experiment. In the unelaborated form the correct response to this task is NVC, but in the elaborated form it is *Some A are C.* In both formats, all three VR systems respond *Some A are C,* and they appear to exhibit an elaboration effect. The problem is that it is impossible to determine whether elaboration has had any effect. Even a theory that assumed that elaborations would have no impact on performance could exhibit an elaboration effect on this task as long as it responded *Some A are C.*

An obvious question is whether there really is an elaboration effect at all or whether it can be completely explained by these kinds of artifacts. Table 5 presents data relevant to this question. These data are the same as those in Table 4 except that the three tasks discussed earlier have not been included. Task 3 has also been excluded because the correct response changes between the unelaborated and elaborated forms (like Task 4). Note that although these tasks have been removed, both the human data and VR's predictions still show an elaboration effect (although it is smaller than before).

VR suggests two explanations for why this effect continues to show up. First, elaborated premises can constrain the annotated

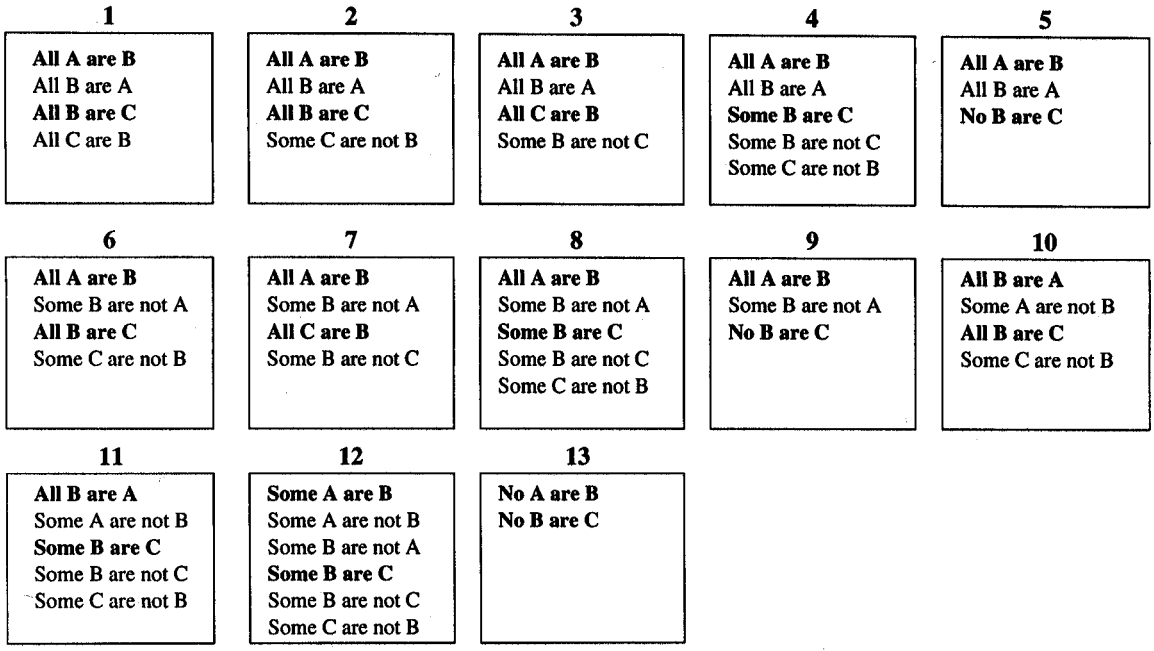

Figure 6. Elaborated and unelaborated (boldface) premises used by Ceraso and Provitera (1971).

model to represent objects that fail to relate the end terms (so that invalid universal conclusions are not generated). Premise elaborations often refer to objects that are not referred to in the premises themselves and that do not associate the end term properties. Because the annotated model is constrained to be consistent with the propositions it encodes (i.e., the propositions are true in the model), it must represent these objects explicitly when given elaborated premises. Because these objects do not relate the end terms, invalid universal conclusions (that would otherwise be generated) are not proposed.

A second reason VR produces an elaboration effect is that elaborations can make certain properties more available (by marking them as identifying), causing conclusions to be generated that might otherwise be overlooked. In some cases, VR will produce an annotated model that relates the end terms but still be unable to generate a legal conclusion because the property that should be the topic does not appear as an identifying property. But if that property appears as the topic of a premise elaboration, it will be marked as identifying, and the appropriate conclusion can be drawn.[6]

Both of these effects can be seen in VR1's behavior on Task 10 in Figure 6. After initial encoding and reencoding (VR1 does none) of the unelaborated premises, *All B are A, All B are C,* VR1 constructs the following annotated model:

$(B'\,A\,C)$

Although this model relates properties A and C, VR1 cannot draw a conclusion because neither of these properties is marked as identifying. Consequently, it responds with NVC. Note that even if property C had, in fact, been an identifying property, VR1 would have responded with *All C are A,* which is still incorrect (this is the response given by VR3 to this premise pair). In contrast, if the original premises had been elaborated to include the additional premises *Some A are not B* and *Some C are not B,* then the annotated model would have been

(A')
$(A'-B)$
$(B'\,A\,C)$
(C')
$(C'-B)$

The premise elaborations have caused two significant changes to the annotated model: (a) Properties A and C now appear as

Table 5
Percentage Correct for Elaborated and Unelaborated Premises in Humans and VR With Four Tasks Excluded

Premise type	Humans	VR1–3 average	VR1	VR2	VR3
Unelaborated	70	78	78	89	67
Elaborated	80	96	100	100	89

Note. VR = verbal reasoning model.

[6] There are at least two other factors that could also lead to elaboration effects but that VR has not simulated. First, premise elaborations could improve performance by blocking the extraction of contradictory indirect knowledge. Second, elaborated premises could influence whether existing model objects are augmented or new objects are created. For example, VR would predict that using elaborated particular premises such as *Some* other *B are C* instead of just *Some B are C* could significantly improve performance by helping participants avoid the fallacy of the undistributed middle term (see earlier discussion).

identifying properties, and (b) additional objects that do not relate the end terms have been included. As a result, VR1 produces the valid conclusions *Some A are C* and *Some C are A* (and only those conclusions). *Some A are C* is one of the four legal alternatives, and so it is selected.

The main point of all of these analyses is that simple but different VR systems all produce the seven major regularities. This result demonstrates that the verbal reasoning hypothesis can account for the major phenomena of syllogistic reasoning (although, as discussed, belief bias effects depend on additional assumptions).

Other predictions. VR makes a number of other predictions about syllogistic reasoning that are not included in the list of standard regularities. We now consider some of VR's major assumptions and derive predictions from each.

First, consider initial encoding. As we have stressed, we assume that initial encoding delivers an annotated model that is consistent with the premises. This assumption predicts that erroneous non-NVC responses are much more likely to be consistent with the premises than would be expected by chance alone. Equivalently, non-NVC errors are less likely to contradict an entailment of the premises than would be expected at random. To test this prediction, we analyzed all 2,424 legal non-NVC errors that occurred in the six data sets presented in Table 1. Of these, only 14 (0.6%) contradicted an entailment of the premises. At random, one would expect 251 (10%) of these non-NVC errors to be inconsistent with the premises.

Initial encoding also assumes that direct knowledge (about a topic) is more available than indirect knowledge. Specifically, during initial encoding only direct knowledge is extracted, and even during reencoding, indirect knowledge is not guaranteed to be available (e.g., VR1 is never able to extract indirect knowledge). This assumption leads to a number of predictions about the *xy–zy* figure. In this figure, the topic of one premise is not mentioned in the other premise. Because direct knowledge is always about the topic of a proposition, this implies that, without any indirect knowledge, the annotated models in the *xy–zy* figure will never relate the end terms. Instead, the first premise will create objects with property X, and the second will create new objects with property Z. Only through indirect knowledge about Y can the end terms be related. In contrast, indirect knowledge is usually not necessary to relate the end terms in the *xy–yz* and *yx–zy* figures (we discuss the *yx–yz* figure subsequently). This leads to a number of predictions. First, there will be more NVC responses in the *xy–zy* figure than in either the *xy–yz* or *yx–zy* figure. Because indirect knowledge is required in the *xy–yz* figure but not the other two, the annotated model is less likely to relate the end terms, and more NVC responses should be observed. Second, participants should be more accurate on indeterminate premise pairs (those without a valid conclusion) in the *xy–zy* figure than on those in either the *xy–yz* or *yx–zy* figure. If NVC responses are more common, then performance should improve on tasks for which this is the correct response. Conversely, participants should be less accurate on determinate premise pairs (those with a valid conclusion) in the *xy–zy* figure than in either the *xy–yz* or *yx–zy* figure.

All but one of these predictions are confirmed in the six data sets from Table 1. NVC responses constitute 55% of legal responses in the *xy–zy* figure, in comparison with only 33% and 41% in the *xy–yz* and *yx–zy* figures, respectively. Sixty-seven percent of legal responses to indeterminate premise pairs in the *xy–zy* figure are correct, but only 45% and 52% are correct in the *xy–yz* and *yx–zy* figures. Conversely, only 45% of legal responses to determinate premise pairs are correct in the *xy–zy* figure, in comparison with 55% in the *xy–yz* figure. Contrary to our prediction, only 46% of legal responses to determinate premise pairs are correct in the *yx–zy* figure, and this is comparable to the 45% in the *xy–zy* figure. There are two determinate tasks in the *yx–zy* figure that are much harder than any of the others (*All B are A, No C are B,* which no participants solve correctly, and *Some B are A, No C are B,* which only 10% of participants solve correctly). For both, the valid conclusion is *Some A are not C,* which goes against the figural effect. It could be that the predicted effect is present but that it is being obscured by the figural effect on these two tasks (recall that the figural effect does not apply to the *xy–zy* figure). Consistent with this view, if these two tasks are excluded from the analysis, the percentage correct on determinate *yx–zy* tasks jumps to 67% (as compared with 45% in the *xy–zy* figure). Even if the two hardest determinate tasks in the *xy–zy* figure are excluded, the percentage correct increases only to 56%, which is still significantly lower.

The assumptions underlying VR's conclusion generation make similar predictions in the *yx–yz* figure. Recall that identifying properties are tried first during generation. But in the *yx–yz* figure, neither end term appears as the topic of a premise, so they will not be marked as identifying after initial encoding. In this figure, then, indirect knowledge is required for an end term to become identifying. Because this is not the case in the *xy–yz* and *yx–zy* figures, VR makes the same predictions about the relationship between *yx–yz* and *xy–yz*/*yx–zy* tasks as it does for the *xy–zy* figure: There will be more NVC responses, indeterminate tasks will be easier, and determinate tasks will be harder. All of these predictions are confirmed in the human data from Table 1. In the *yx–yz* figure, NVC responses constitute 55% of the legal responses (in comparison with 33% and 41% in the *xy–yz* and *yx–zy* figures), 80% of legal responses to indeterminate tasks (in comparison with 45% and 52% in the other figures), and only 36% of legal responses to determinate tasks (in comparison with 55% and 46% in the others).

Turning to reencoding, VR assumes that whenever the annotated model fails to lead to a legal conclusion, participants reencode the premises in an effort to extract more knowledge and augment their model. If participants are constrained to respond within a short period of time, they may be forced to give up before they have completed all of their reencoding attempts. Consequently, they may miss conclusions that they would produce with more time. This predicts a higher proportion of NVC responses from timed participants in comparison with the same participants given unlimited time (at least when generating conclusions rather than evaluating them). The data sets labeled *timed* and *revised* in Table 1 provide a relevant comparison. In the timed experiment, participants received all 64 premise pairs and were asked to respond within 10 s. Afterward, they were presented with the same tasks, along with the responses they had just provided and were given 1 min to revise their answers if desired. Consistent with VR's prediction, NVC responses

Table 6
Other Predictions From VR and Their Empirical Status

Prediction	Empirical status
Non-NVC errors more likely to be consistent with premises than expected by chance	Confirmed
More NVC responses in *xy–zy* figure than in *xy–yz* figure	Confirmed
More NVC responses in *xy–zy* figure than in *yx–zy* figure	Confirmed
More correct responses to indeterminate *xy–zy* tasks than to indeterminate *xy–yz* tasks	Confirmed
More correct responses to indeterminate *xy–zy* tasks than to indeterminate *yx–zy* tasks	Confirmed
Fewer correct responses to determine *xy–zy* tasks than to determinate *xy–yz* tasks	Confirmed
Fewer correct responses to determinate *xy–zy* tasks than to determinate *yx–zy* tasks	Marginally confirmed
More NVC responses in *yx–yz* figure than in *xy–yz* figure	Confirmed
More NVC responses in *yx–yz* figure than in *yx–zy* figure	Confirmed
More correct responses to indeterminate *yx–yz* tasks than to indeterminate *yx–zy* tasks	Confirmed
More correct responses to indeterminate *yx–yz* tasks than to indeterminate *yx–zy* tasks	Confirmed
Fewer correct responses to determinate *yx–yz* tasks than to determinate *yx–zy* tasks	Confirmed
Fewer correct responses to determinate *yx–yz* tasks than to determinate *yx–zy* tasks	Confirmed
More NVC responses under time constraints	Marginally confirmed

Note. NVC = no valid conclusion; VR = verbal reasoning model.

constituted 56% of legal responses in the timed condition in comparison with 47% in the revised condition. It is possible that providing participants with their previous responses could account for this effect if they considered NVC responses to be more tentative than others and tried harder to revise them (in VR, NVC is a response to giving up, and so it may be more tentative). This interpretation becomes less likely, however, when one considers the other data sets from Table 1. In three of the four (unlimited, Week 1, and Week 2), the percentage of NVC responses was below the 56% observed in the timed data (34%, 45%, and 48%). And although two of these studies used completely different participants, the first (unlimited, in which only 34% of responses were NVC) used participants from the same pool as the timed experiment (students at the University of Milan). The one study that did not show a smaller percentage of NVC responses (Inder [1986, 1987], with 56%) involved only 3 participants. Thus, although the evidence is not conclusive, the data seem to confirm VR's prediction that time constraints should lead to more NVC responses. Table 6 summarizes the preceding predictions, along with their empirical status.

Individual Data

Data from individual participants provide an even more stringent test for a theory than do aggregate data. Nevertheless, essentially no work in the reasoning literature has attempted to analyze and predict such data. A computational system like VR, however, provides a natural approach to attacking this issue. The first step is to identify the aspects of VR that could most plausibly differ across participants and to formulate those aspects as a set of explicit parameters. Then VR can be run with different parameter settings to tailor its performance to that of individual participants. Such an analysis can provide insight into the ability of the verbal reasoning hypothesis to account for individual data.

Individual-differences parameters for VR. Figure 7 presents such a set of individual-differences parameters. The first six parameters affect how VR encodes propositions. The next three (7–9) influence VR's conclusion generation process. Parameters 10 through 21 control what indirect knowledge VR extracts during reencoding. Finally, Parameter 22 controls whether or not VR attempts to falsify its putative conclusion by searching for alternative models (in keeping with mental model theory; Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991). According to the verbal reasoning hypothesis, such a falsification strategy should be much less important in explaining behavior than are the verbal processes or encoding and reencoding. We included the parameter to allow us to test whether searching for alternative models would help VR's fits. Appendix C describes the details of how these parameters were implemented.

To reduce the size of this parameter space, we performed some preliminary analyses to identify parameter values that were most and least important in fitting the individual data. We divided the 64 premise pairs into four sets of 16 tasks such that each set contained the same number of tasks from each figure and of each premise type[7] and randomly selected 20 participants from the 103 in Table 1.[8] We then computed the best-fit parameter settings for each participant on all four task subsets, that is, the parameter settings that caused VR to produce the most correct predictions on those tasks. This and all subsequent fits were computed with the ASPM system, a suite of computational tools that makes it possible to fit and analyze symbolic parameter models with as many as 10 billion parameter settings (Polk, Newell, & VanLehn, 1995; Polk, VanLehn, & Kalp, 1995). On the basis of these results, we computed two subsets of parameter space, one large and one small. The large space excluded parameter values that the preceding analysis suggested could be safely removed and whose absence we intuitively thought would not be important. Those values are indicated by a minus sign in Figure 7. The analysis also suggested removing

[7] The specific tasks in each set were as follows (in terms of the notation from the middle of Figure 1): (a) {ObaEcb, EbaEcb, AbaOcb, AbaIcb, ObaEbc, ObaAbc, IbaObc, AbaEbc, EabIcb, IabAcb, AabEcb, AabAcb, OabObc, OabIbc, IabEbc, IabIbc}, (b) {AabIbc, EabOcb, EbaAbc, ObaAcb, IabOcb, IbaEbc, EabEbc, ObaOcb, IabIcb, AbaAbc, ObaIcb, OabAbc, AabIcb, EbaOcb, EbaIbc, EabAbc}, (c) {ObaIbc, OabAcb, AbaEcb, EabIbc, IbaAcb, EabObc, IbaIbc, EabEcb, ObaObc, AbaAcb, IabObc, OabEcb, IabAbc, EbaIcb, EabAcb, AbaObc}, and (d) {AabEbc, Oableb, EbaObc, IbaEcb, IbaAbc, AabObc, OabOcb, IbaIcb, EbaEbc, AabAbc, IbaOcb, IabEcb, AbaIbc, OabEbc, AabOcb, EbaAcb}.

[8] Participants 2 and 11 from unlimited; 4, 13, and 18 from timed; 3, 5, 6, 11, 15, and 16 from revised; 5, 7, 8, and 13 from Week 1; 2, 6, 13, and 20 from Week 2; and 1 from Inder (1986, 1987).

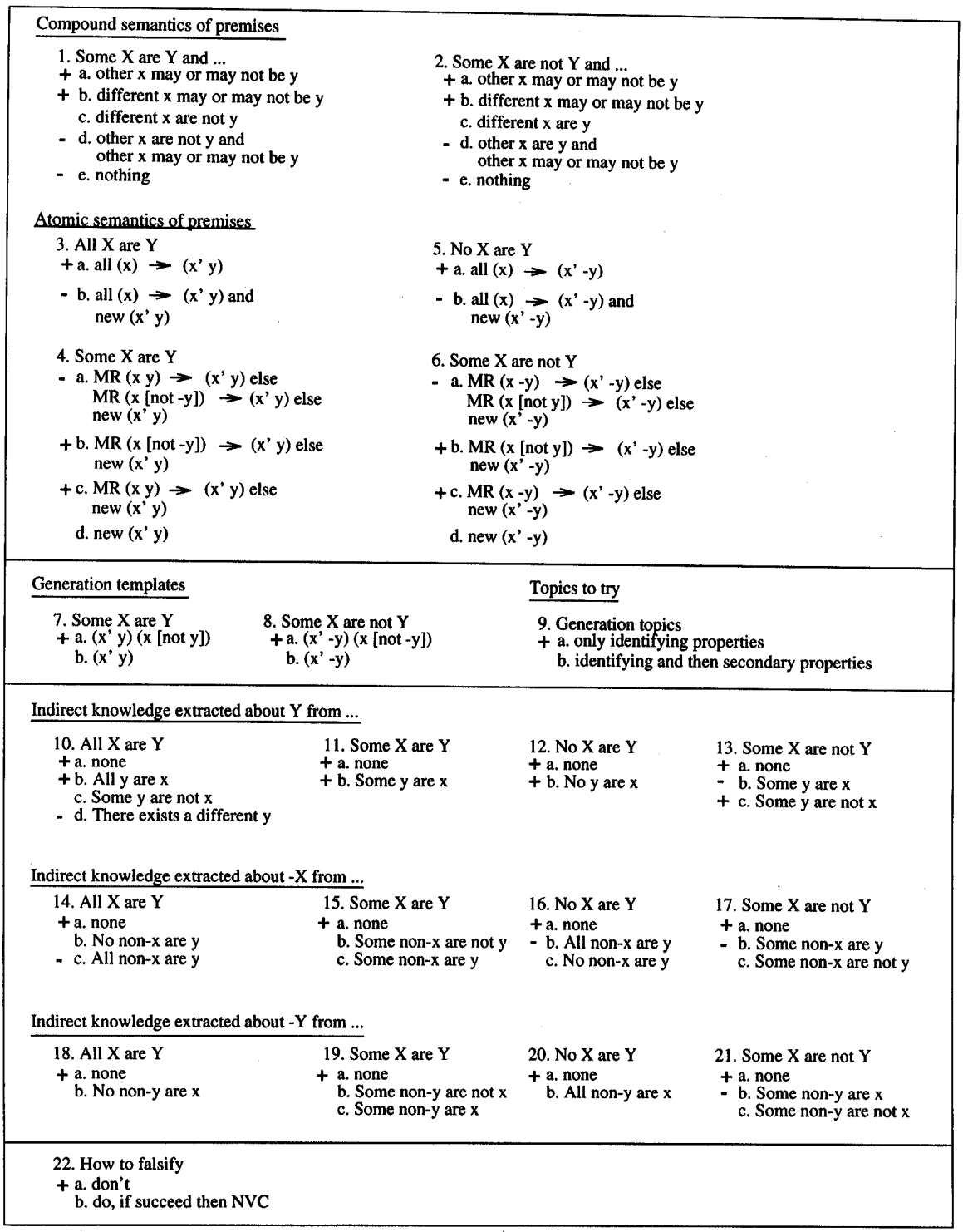

Figure 7. Individual-differences parameters for VR. + = most useful values; − = least useful values; MR = most recently accessed, ' = identifying property.

Value b for Parameter 22, which causes VR to attempt to falsify its putative conclusions by searching for alternative models (only 2% of our best-fitting settings contained that value). But because falsification is central to other theories of deduction, we wanted to investigate it further and did not remove it. The small space includes only parameter values that were critical in achieving good fits for the 20 participants in the preceding analysis. These values are indicated by a plus sign in Figure 7.

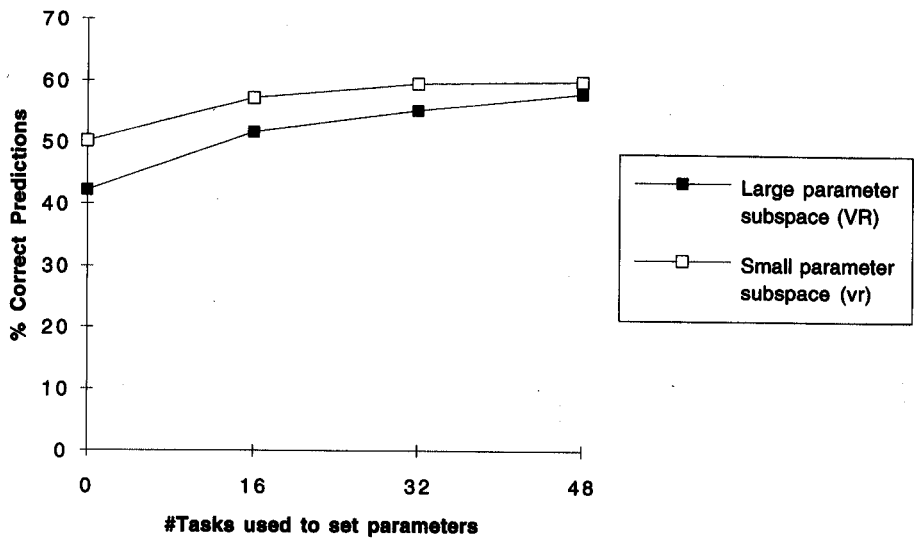

Figure 8. VR's fits to individual data.

In the course of these analyses, we noticed that VR was fitting tasks that did not involve *All* much better than it was fitting those that did. Specifically, VR responded with NVC to these tasks more often than did the participants. This phenomenon is reminiscent of Lee's (1983) *A-effect:* Premise pairs involving *All* seem to facilitate the generation of conclusions. Inder (1987) argued that the effect could be explained in terms of a syntactic generation strategy for premise pairs involving *All* (Ford, 1995, recently made a similar proposal). The basic idea is that, given the premise *All y are x*, participants may simply respond with the other premise after having replaced the *y* in it with *x*. For example, given the premise pair *All y are x, Some z are not y*, this strategy would replace the *y* in the second premise with *x* and produce the response *Some z are not x*. Note that this strategy applies only if the middle term (*y* in this case) appears as the topic of an *All* premise. A more extreme version of this strategy would perform the replacement even if the middle term appeared only as the predicate of an *All* premise. We augmented the small space with a parameter that allowed for both strategies, as well the absence of either:

23. All substitution strategy
a. do not apply it
b. apply if middle term appears as topic of "All"
c. apply for any "All" premise

VR's fits to individual data. Figure 8 presents VR's average fits to the 103 participants for both the large parameter subspace (without the all-substitution strategy) and the small parameter subspace (with it). We used complementary tasks for setting the parameters and evaluating the fit. Specifically, we chose four task subsets that contained 16 tasks, four that contained 32 tasks, and four that contained 48 tasks. Each of the 64 tasks was equally represented across the subsets of tasks. This is important because it means that all 64 tasks were also equally weighted in the evaluation (described later).[9] We also included the null set (no tasks). Then, for each participant, we computed all of the best-fitting parameter settings for each subset of tasks. In the case of the null set, no parameter settings could fit better than any other, so all were included.

Given such a set of best-fitting settings, we then computed their average fit (i.e., the average proportion of tasks for which VR and the participant gave the same response) over all of the other tasks (those that were not used to set the parameters) for which the participant gave a legal response. If VR predicted N responses for a task, one of which was the observed response, then the fit for that task would be $1/N$ (usually N was 1). The overall fit was the sum of these task fits divided by the number of tasks. For the large parameter space, it was not feasible to average over all best-fitting parameter settings in the no-task case because this constituted searching the entire space. Instead, we took a random sample of 1,000 parameter settings and averaged their fits on the 64 tasks. Figure 8 presents the average of all of these mean fits for each space and set of fitting tasks.

One interesting aspect of these results is that the smaller parameter space performs better than the larger space. In the four sets of fitting tasks in Figure 8, the smaller space produces fits that are between 2% (48 fitting tasks) and 8% (no fitting tasks) better than the larger space. We believe that two main factors contributed to this unexpected result. First, the small space, but not the large space, included the all-substitution strategy. Consequently, the small space had an advantage in fitting participants who applied this strategy despite the greater flexibility of the large space. Second, the added flexibility of the larger space could reduce the fit's quality because of overfitting (by including parameter settings that lead to rarely observed behaviors). For example, suppose we were using 16 tasks to set the parameters and the optimal fit on those tasks was 15 correct predictions.

[9] The sets of 16 tasks are specified in footnote 7. The sets of 32 tasks were formed by combining pairs of those 16 tasks (specifically, Sets 1 and 2, 3 and 4, 1 and 4, and 2 and 3). Similarly, the sets of 48 tasks were formed by combining sets: {2, 3, 4}, {1, 3, 4}, {1, 2, 4}, and {1, 2, 3}.

Then all and only the settings that produced that level of fit would be included in the analysis, regardless of how common or uncommon we might consider those settings a priori. Thus, plausible settings that produce a slightly lower fit (e.g., 14 of 16) would not be included, but implausible settings that just happened to do well on the set of fitting tasks would be. These implausible settings would presumably not do very well in fitting the evaluation tasks, and the average fit would go down. This problem is not nearly as severe for the smaller space because it includes only plausible settings. Furthermore, adding fitting tasks alleviates the problem because it becomes less and less likely that the implausible settings will fortuitously produce optimal fits on the fitting tasks. This explains why the disparity in fits between the large and small spaces is largest in the no-task case (in which none of the implausible settings are ruled out in the fit) and smallest in the 48-task case (in which almost all of them are).

As expected, the fits improve as additional tasks are used to set the parameters. When no tasks are used, the fit indicates only how well the settings in the space do on average. As more fitting tasks are added, VR can fine-tune its behavior to model the individual participants more closely. Also note that the first few fitting tasks lead to much larger improvements than do the last few; the difference between using 0 and 16 tasks is much larger than the difference between using 32 and 48. Apparently, the first few tasks rule out portions of parameter space that are way off the mark, whereas the last few can lead only to small refinements in the predictions, not major improvements. Indeed, the smaller space appears to level off at approximately 60% correct predictions. Adding fitting tasks would probably not help much. Thus, 60% correct predictions is a stable estimate of how well the smaller space can do in predicting individual behavior. It is important to note that, in one sense, these are zero-parameter fits; none of the data used to evaluate the fits contributed to the setting of the parameters (because different tasks were used for each purpose). Consequently, the fit qualities are not artificially high as a result of overfitting.

Comparisons with reference theories. Figure 9 presents several additional analyses that provide useful reference theories for comparison. On the far left is the expected percentage of accurate predictions for a completely random theory. Because there are nine legal conclusions, the probability of such a theory predicting any particular response is 11% (1 in 9). We also computed the number of identical responses between pairs of participants and averaged over all such pairs. On average, 44% of responses were consistent across pairs of participants (other-subjects in Figure 9). The correct theory predicts correct performance on every task. For tasks in which there are multiple correct responses, this theory chooses randomly. Such a theory predicts 49% of the observed responses from the individuals in our data sets. Test–retest is the average proportion of responses given by a specific participant during one session that exactly match his or her responses to the same tasks a week later (on average, 58% of responses are identical). This is based on the one data set (Johnson-Laird & Steedman, 1978) that retested the same 20 participants a week later. VR's average fits to this same subset of participants are labeled in Figure 9 as Test–retest (VR) (59%; the large parameter space) and Test–retest (vr) (62%; the small parameter space). The modal case (59%) is a post hoc analysis in which the most common response over all of the participants is treated as the prediction. Finally, VR and vr correspond to the model's average fit to all 103 participants for the large and small parameter spaces, respectively. All of VR's fits in this figure were computed using the four sets of 48 tasks to set the parameters and the other tasks to evaluate the fit.

The reference theories in Figure 9 provide a baseline against which to compare VR's fits. As one would expect, VR's predictions are significantly better than those that would be produced by predicting randomly or assuming that responses were always correct. What is much more impressive is that VR's fits are at or beyond the test–retest reliability of the participants themselves. Test–retest reliability gives a pessimistic estimate of the stability and systematicity of the participants and thus provides an estimate of how well a fixed deterministic theory could possibly do without overfitting the data. (The estimate is pessimistic because the two tests occurred a week apart. If they had been administered together, the test–retest reliability estimate would probably be a little higher, although such a measure would also confound reliability with memory of the responses.) Consequently, one could hardly ask for more from VR without attempting to capture the instability of the participants' performance over time (e.g., learning). Modeling instability in participant behavior is a natural next step for our approach but one that we have not yet undertaken.

The other-subjects and modal cases in Figure 9 provide insight into how much of VR's fit could be due to capturing group trends versus individual differences. The other-subjects result gives an estimate of how well a theory without any parameters could do if it had access only to 1 participant's data. Such a theory could be tailored to predict that 1 participant perfectly, but, without any parameters to allow for individual differences, only 44% of its predictions would be accurate on average. Clearly, VR is capturing significantly more of the behavior than this.

The best that any fixed theory (i.e., one without parameters) could possibly do would be to behave like the modal theory (i.e., to predict the most commonly observed response for every task). This theory's success (59% correct predictions) demonstrates what a large proportion of behavior could, in principle, be explained without reference to any individual differences. Of course, the theory provides an upper bound that is unlikely to be attained in practice. Nevertheless, it is worth pointing out that the modal theory's accuracy is comparable to the test–retest reliability of the participants. That is, the most common response is as good a predictor of performance as is a specific participant's own behavior a week earlier. In any case, the fact that VR's predictions are as good or better than the modal theory shows that VR can predict behavior more accurately than can any fixed theory.

It is also interesting to compare the modal theory's performance with VR when the number of tasks used to set the parameters is zero (VR is being treated as a fixed theory or a set of fixed theories). In the case of the small parameter space (which was designed to include only the most common parameter settings), the quality of VR's predictions is within 10% of optimality.

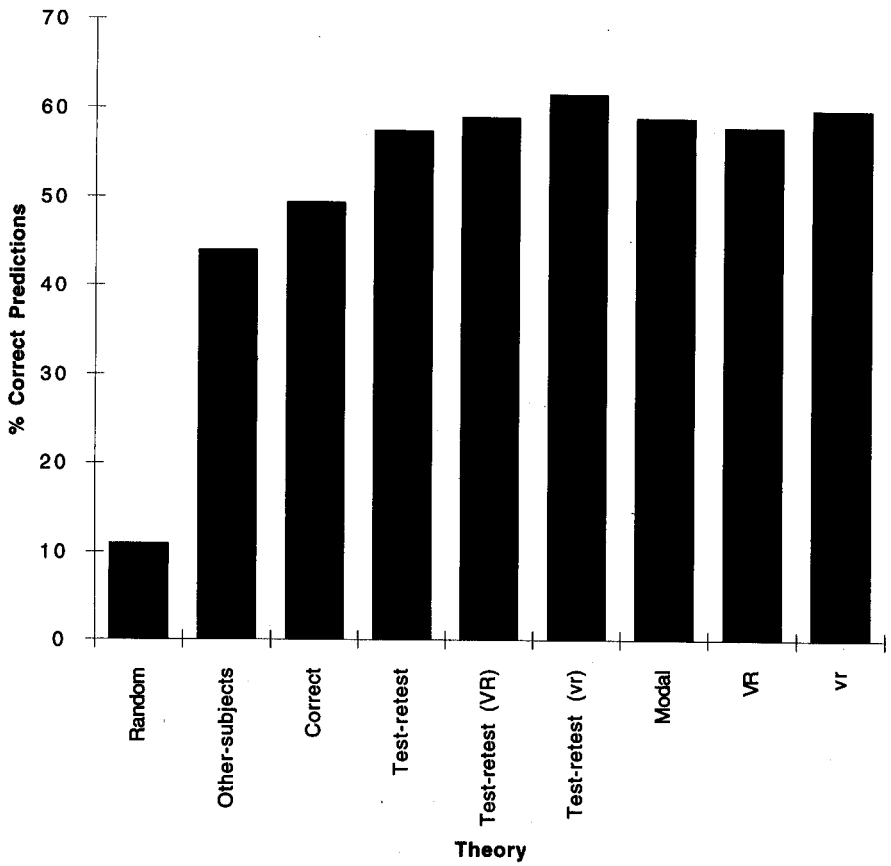

Figure 9. Comparison of VR's fits with various reference theories. VR = large parameter space; vr = small parameter space.

Because the smaller space produces such good fits, the few parameters that it allows to vary account for many of the individual differences. One can identify these parameters in Figure 7 because they have more than one plus sign beside them (the others have a fixed value and cannot vary). These parameters can be grouped into two major classes: those involving the interpretation of particular premises (*Some* and *Some not* [Parameters 1, 2, 4, and 6]) and those involving the extraction of indirect knowledge about a proposition's predicate (Parameters 10–13). In addition to these two, the all-substitution strategy (the new parameter) is another important source of individual differences; including more than one of its values (especially Values a and c) was important in obtaining good fits for different participants. This result agrees with data from Ford (1995) indicating that the use of a syntactic substitution strategy is an important source of individual differences. According to VR, then, the three most important sources of individual differences in syllogistic reasoning are (a) the interpretation of the quantifier *Some,* (b) what indirect knowledge participants can extract from propositions about their predicates, and (c) whether or not participants apply the all-substitution strategy. Of course, the crucial point from this analysis is that the verbal reasoning hypothesis, as manifested in VR, can and does account for the detailed behavior of individual participants solving categorical syllogisms.

VR's fits to artificial data. To evaluate the range of behaviors admitted by VR, we computed its fits to some artificially generated data. These fits are presented in Figure 10. For random performance, we generated five data sets consisting of randomly generated responses to all 64 tasks. The fits presented are the averages over these five data sets for each parameter space. Correct performance consisted of correct responses to all of the tasks. If a task had more than one correct response, predicting any of them was considered an accurate prediction. Modal performance corresponded to the behavior of the modal theory described earlier; for each task, we selected the most frequently observed response. Because the point of these analyses was to determine to what extent each type of behavior could be simulated by some parameter setting, we computed the true optimal fits over all 64 tasks rather than over a subset of tasks (as earlier).

Figure 10 demonstrates some interesting points about the scope of VR, that is, what types of behavior it can and cannot fit well. First, note that with the large parameter space, VR is able to fit correct performance perfectly. In other words, there is at least one setting (actually there are many) in parameter

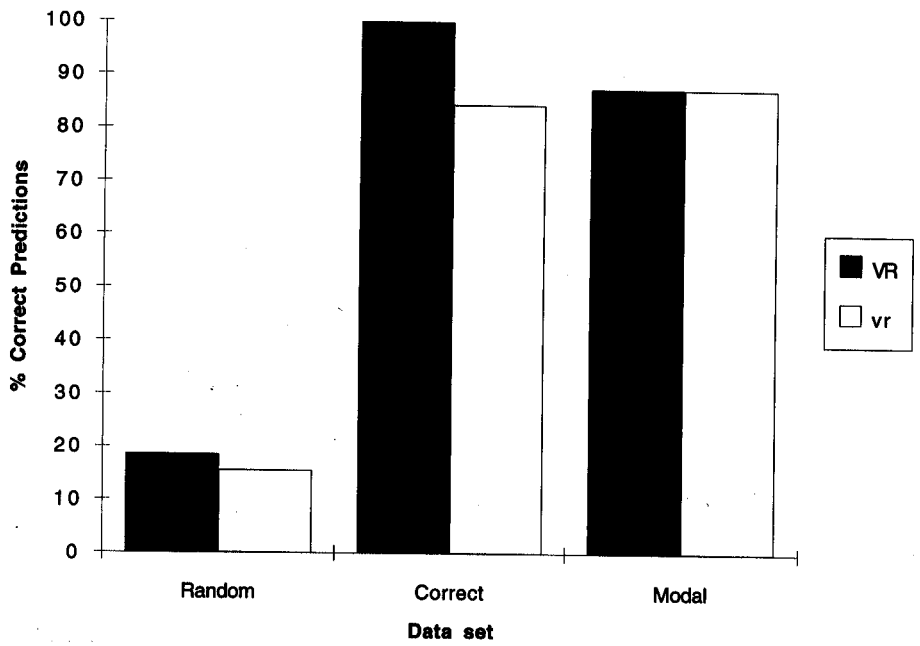

Figure 10. VR's fits to artificially generated data. VR = large parameter space; vr = small parameter space.

space that leads VR to solve all of the tasks correctly (VR allows for perfect performance). Many of these parameter settings did not involve falsification. This fact is important because it shows that verbal reasoning does not necessarily give rise to poor performance. A participant could do well even without relying on sophisticated reasoning-specific mechanisms.

Second, VR can fit modal data very well but not perfectly. Perhaps if VR was a perfect theory it would be able to predict modal data completely. On the other hand, it is possible that no participants completely exhibit modal performance, and trying to predict it could lead to incorrect theories (for a compelling discussion of some of the dangers of trying to fit aggregate rather than individual data, see Siegler, 1987).

Third, VR is very bad at fitting random data. Less than 20% of its predictions are accurate on average, and this is very close to what would be expected even if the parameters were not being fitted at all (11%). The slight improvement in fit can be attributed to overfitting; the data being used to set the parameters are also being used to evaluate the fit (recall that this was not a problem in Figure 8). The fact that VR cannot fit random data demonstrates its falsifiability. It is possible to observe data with which VR is inconsistent.

Analysis of falsification. Finally, Figure 11 presents an analysis of how well the small parameter space (vr) can fit individual data with and without falsification. (It was not possible to obtain these fits for the large parameter space for computational reasons. See the discussion of loosely coupled tasks in Polk et al. [1995] for a detailed treatment of the problem we encountered.) The top row in each pair presents the percentage of legal responses that were accurately predicted when falsification was not used (Parameter 22 had Value a), and the bottom row presents the accuracy when falsification was used. All 64 tasks were used to set the parameters in both cases. The boxes indicate which fit was better. Heavy boxes indicate a disparity in the fit of more than 5%. The six pairs of rows correspond to the six data sets from Table 1. The first five involved 20 participants each, and the last (Inder, 1986, 1987) involved only 3 participants.

Two additional values were added to the falsification parameter so that it consisted of four values:

22. How to falsify
a. don't
b. do, if succeed then NVC
c. do, if succeed then respond based on new annotated model
d. do, recursively apply (c) above

The last two values correspond to falsification strategies similar to that proposed by Johnson-Laird (Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991). With Value c, VR attempts to construct an annotated model in which the premises are still true but the conclusion is not. If it succeeds, it generates a new conclusion based on the resulting model (rather than responding NVC, as Value b does). With Value d, VR recursively applies this strategy. That is, every time it generates a putative conclusion, it tries to construct an alternative model. If it succeeds and that new model supports yet another (different) conclusion, then VR attempts to falsify it. It continues in this way until it cannot falsify a conclusion or until the only conclusions it can generate have already been falsified (in which case it responds NVC).

The results in Figure 11 illustrate a number of points about falsification. First, there is not a very large difference in the fits with and without falsification. Only 15 participants' fits (out of 103) were affected by more than 5% on the basis of this param-

	Subject																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
a	48.9	54.7	46.7	67.3	52.5	68.3	58.1	60.9	66.7	58.1	60.0	50.0	50.9	54.7	31.7	75.0	58.7	82.5	54.7	65.0
bcd	60.0	50.0	41.6	59.2	45.9	61.9	61.3	57.8	60.3	56.5	56.7	48.4	41.5	50.0	36.7	71.9	55.6	81.0	56.3	61.7
a	61.3	64.5	69.8	80.6	85.9	74.2	76.6	79.7	48.4	57.8	66.7	64.5	38.7	63.3	80.0	59.4	65.6	84.4	68.8	49.2
bcd	62.9	56.5	66.7	77.4	84.4	75.8	75.0	82.8	45.3	57.8	65.1	56.5	41.9	61.7	73.3	62.5	65.6	81.2	62.5	46.0
a	66.7	62.5	82.8	79.4	81.3	66.7	59.4	87.5	43.8	70.3	61.9	67.2	30.2	42.1	68.3	68.3	85.9	85.9	67.2	48.4
bcd	66.7	60.9	79.7	76.2	79.7	68.3	56.3	87.5	42.2	67.2	61.9	62.5	28.6	38.6	66.7	66.7	82.8	82.8	62.5	46.8
a	55.7	63.5	63.5	50.0	81.3	77.4	64.5	73.4	75.0	60.9	68.8	72.6	62.5	65.1	79.7	54.2	61.9	47.5	76.6	62.9
bcd	55.7	58.7	55.6	46.8	81.3	75.8	59.7	75.0	76.6	60.9	68.8	71.0	62.5	58.7	68.8	54.2	58.7	44.3	75.0	66.1
a	66.7	81.3	67.2	56.3	77.8	77.8	69.4	78.1	75.0	75.0	78.1	67.2	72.6	79.7	76.6	58.7	64.5	61.3	77.4	60.9
bcd	69.8	81.3	65.6	54.7	79.4	79.4	69.4	78.1	71.9	73.4	79.7	65.6	72.6	79.7	73.4	57.1	64.5	54.8	77.4	64.1
a	73.0	68.9	57.8																	
bcd	71.4	70.5	57.8																	

Figure 11. Small parameter space (vr) fits with and without falsification (bcd and a, respectively).

eter. Furthermore, in only 1 case out of 103 was the fit with falsification more than 5% better than the fit without. This result is noteworthy given that the parameter space was 3 times larger with falsification than without it (Values b, c, and d were all considered consistent with falsification), allowing for a wider range of behavior.

One possibility is that VR has other parameters that do most of the work of falsification. That seems unlikely. Most of the parameters simply allow for obvious alternative interpretations of the premises (during encoding and reencoding), whereas others vary the conditions under which different conclusions are drawn (by matching against the annotated model). None of them (aside from the falsification parameter itself) are related in any obvious way to falsification. This argument is particularly compelling for the small parameter space. One would have to assume that these few parameters, constructed independent of falsification, just happen to be able to interact in a way that closely models falsification. Furthermore, one must assume that they can do so while still capturing the wide variety of individual differences across participants.

One plausible interpretation is that the situations in which VR produces a putative conclusion that it can then falsify are the exception rather than the rule. Most versions of VR solve approximately 50% of syllogisms correctly, and these conclusions would not be falsified even if VR tried. Furthermore, many of VR's errors involve NVC responses to a syllogism with a valid conclusion. Here, again, the process of constructing an alternative model to falsify the putative conclusion will not be relevant. Thus, for most syllogisms, falsification will not have an impact one way or the other on VR's predictions. This fact is reflected in the falsification analysis: For 88 of 103 participants, the falsification parameter did not make a difference either way. For 14 of the other 15 participants, however, falsification decreased the quality of the fit. The natural conclusion is that falsification is an unnecessary and somewhat harmful assumption for VR. Of course, this analysis does not rule out falsification, and it may still be an important assumption in other theories (e.g., mental model theory). The point is that syllogism data can be accurately explained without positing a falsification strategy, in keeping with the hypothesis that such a reasoning-specific strategy is less important than verbal processes in explaining deduction.

Discussion

VR's behavior exemplifies verbal reasoning. After initial encoding, VR repeatedly reencodes the premises until it can generate a legal conclusion (see Figure 2). And all of these processes (initial encoding, reencoding, and conclusion generation) are fundamentally linguistic; that is, they transform verbal information into a semantic representation and back again. Far from being devoted exclusively to deduction, their main role is presumably language processing. Other tasks (e.g., reading) might also require giving up on reencoding to avoid excessive delays, although one could argue that VR's exhaustive reencoding would be unique to reasoning. In any case, giving up is certainly much less central to VR's behavior than are reencoding and conclusion generation. Falsification is the one process we explored that is clearly designed specifically for reasoning. Its purpose is to ensure deductive validity, and so it would presumably not be relevant in other domains. And unlike linguistic processes, it does not map between verbal and semantic representations; rather, it transforms one semantic representation into another. If used, falsification could be as important as the other processes, especially if a sequence of putative conclusions is successively falsified, leading to a completely different response. As mentioned earlier, however, falsification proved to be an unnecessary assumption in achieving the fits we did with VR. Only 1 participant's fit (out of 103) improved significantly when VR falsified as opposed to when it did not.

Although VR is based on the verbal reasoning hypothesis, it builds on ideas from both model-based and rule-based theories of deduction. It uses an annotated model as its primary data structure, and the processes of encoding and reencoding resemble the application of formal inference rules. It is not really a hybrid approach, however. The most important assumption of the verbal reasoning hypothesis—and the one that distinguishes it from both mental model theory and rule-based accounts—is that the central processes in deduction are borrowed from language comprehension and generation. Although these processes resemble the application of formal inference rules, the critical point is that they are actually linguistic processes that map between verbal and semantic representations rather than general-purpose inference rules that transform one semantic representation into another. According to the verbal reasoning hypothesis, one need not posit a mental logic (at least for this task) because linguistic processes are sufficient.

VR's success—in producing the major phenomena, in making accurate novel predictions, and especially in modeling the detailed behavior of individual participants—demonstrates that verbal reasoning provides an accurate and detailed account of the behavior of untrained participants on categorical syllogisms. A natural question is how well previous theories have fared in comparison with verbal reasoning.

Woodworth and Sells (1935) were two of the first psychologists to attempt to understand human behavior on categorical syllogisms. They proposed a theory, known as the atmosphere hypothesis, to account for a subset of errors that people make. Begg and Denny (1969) provided a concise characterization of this hypothesis:

1. If either premise is particular (*Some* or *Some not*), then the conclusion will tend to be particular. Otherwise, the conclusion will tend to be universal (*All* or *No*).

2. If either premise is negative, then the conclusion will tend to be negative. Otherwise, the conclusion will tend to be positive.

The basic idea behind this hypothesis is that the premises produce a "set" in the participant reading them and that this set biases the conclusions produced. As might be expected for the first theory of syllogistic reasoning, the atmosphere hypothesis is not as complete or accurate as more recent theories, including VR. For one thing, it was intended only to be a theory of errors. As such, it specifies not how people solve syllogisms correctly (at least for syllogisms whose correct conclusion is not atmospheric) but only how they solve them incorrectly. Furthermore, it cannot account for NVC responses, which often constitute a significant portion of those observed. It also fails to predict the figural, belief bias, and elaboration effects. VR, on the other hand, provides a complete process model for both correct and incorrect responses (whether NVC or not) and computationally models all of the major regularities.

Chapman and Chapman (1959) argued that many errors could be explained by assuming that participants had illicitly converted one (or both) of the premises (assuming that *All x are y* leads to *All y are x* and that *Some x are not y* leads to *Some y are not x*). Revlis (1975b) presented a detailed computational model based on the same assumption. Both of these theories account for a number of errors but fail to provide a process model of correct performance; they assume the existence of what Revlis called "an unspecified deduction operation." Empirically, neither theory predicts the belief bias effect. A more serious problem is that, by assuming that the premises are converted, these theories predict that a syllogism's figure should not affect behavior (the only difference between figures is the order of terms, and converting the premises neutralizes that difference). Consequently, without some modification, these theories are inconsistent with the figural effect. Again, VR provides a complete process model for both correct and incorrect responses and predicts all of the major regularities, including the figural effect.

Erickson (1974), Guyote and Sternberg (1981), and Fisher (1981) all proposed that participants solve syllogisms by manipulating representations analogous to Euler circles. These theorists assumed that participants encode each premise into Euler circles, combine these representations (sometimes selecting a subset of the initial encodings), and then read out the conclusion from the result. These accounts are relatively complete; they provide detailed explanations of all correct and incorrect answers, as well as NVC responses. But because all terms have equal status in Euler circles the theories do not always predict the figural effect (which depends on the end terms being distinguished in some way). They can occasionally do so by assuming that each conclusion is associated with a specific Euler circle configuration (if the figural conclusion matches the configuration and nonfigural conclusions do not). But in some cases, syllogisms in opposing figures are assumed to produce the same configuration (e.g., *No A are B, No B are C* and *No B are A, No C are B*). Clearly, these theories cannot predict figural effects for both tasks. Furthermore, none of these theories has modeled the belief bias effect. In contrast, VR has been used to model both phenomena.

Mental model theory (Johnson-Laird, 1983; Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991) shares the most similarities with VR, and so we discuss it in some detail. Johnson-Laird and Bara (1984) outlined the model theory of categorical syllogisms in terms of three basic steps.

1. Construct a mental model of the premises (i.e., of the state of affairs they describe).

2. Formulate, if possible, an informative conclusion that is true in all models of the premises that have so far been constructed. An informative conclusion is one that, when possible, interrelates terms not explicitly related in the premises. If no such conclusion can be formulated, then there is no interesting conclusion from syllogistic premises.

3. If the previous step yields a conclusion, try to construct an alternative model of the premises that renders it false. If there is such a model, abandon the conclusion and return to Step 2. If there is no such model, then the conclusion is valid.

The authors implemented the central tenets of this theory in computer programs and predicted that two main factors should affect the difficulty of a syllogism: the figure of the premises and the number of mental models that have to be constructed. In this theory, the figure of the premises affects the difficulty of constructing an initial model and of formulating a conclusion whose subject appeared before its object in the premises. The number of models is important because each model makes demands on working memory. If the models required of a task exceed the capacity of working memory, then responses will be

based on a subset of the models and responses will tend to be erroneous. Consequently, tasks that require only one model are predicted to be easier than those that require more. Johnson-Laird and Bara (1984) presented experiments that confirmed these predictions and that showed that their computer program could account for the most frequent responses on almost all 64 syllogism tasks.

Mental model theory is probably the most successful of any of the previous theories. It explains correct as well as incorrect responses (whether NVC or not) and has been used to explicitly predict three of the major phenomena (difficulty, figural, and belief bias effects). Furthermore, the specific predictions of the theory in regard to the 64 tasks show evidence for three of the other four phenomena (validity, atmosphere, and conversion effects). Finally, although the elaboration effect has not been explicitly modeled, the theory also clearly predicts it, because the theory assumes that the mental model must be consistent with the premises. Consequently, like VR, elaborating the premises to be unambiguous will constrain the mental model to be closer to valid, leading to better performance.

It should be clear that VR was significantly influenced by mental model theory; the annotated models used in VR were based directly on mental models (with some slight modifications). It may be tempting to suppose that VR is really just a variant of this approach and that it does not represent an alternative theory at all. But there is an important difference in how the two theories view the process of syllogistic reasoning. In mental model theory, the heart of the process is the search for alternative models, a process that is devoted primarily to reasoning. It follows the transduction paradigm in assuming that the operations of encoding and decoding serve the purpose of transduction and that deduction itself is carried out by the search for alternative models. In contrast, encoding and reencoding are at the heart of deduction itself, according to the verbal reasoning hypothesis.

VR's reencoding process represents a fundamental departure from the falsification process of model theory. For example, falsification is triggered by the production of a putative conclusion, whereas reencoding in VR is triggered when the system fails to produce a legal conclusion. Also, the goal of falsification is to refute a conclusion, whereas reencoding simply tries to augment the annotated model by extracting more information from the premises. Falsification is a fairly sophisticated strategy in comparison with reencoding, especially for untrained participants. Such participants must appreciate the need to consider as many alternatives as possible, and they must also possess the skills to modify an existing model in such a way as to falsify their putative conclusion without falsifying the premises. In contrast, reencoding is an obvious response to a lack of knowledge, especially when the only available sources of knowledge are verbal premises. Finally, falsification is an explicit attempt to ensure deductive validity, and, as such, it is specific to deduction and would not be appropriate in other tasks. Reencoding, on the other hand, is a general linguistic process that extracts more information from verbal material and is not reasoning specific.

VR produces NVC as a default response when it fails to come up with anything else. Thus, unless VR knows that its other processes are guaranteed to produce any valid conclusions, it cannot be sure that NVC is really the correct answer. The same is presumably true of mental model theory because it falsifies only putative conclusions. How could it know that all of the conclusions it did not falsify are invalid? The fact that VR and model theory cannot be sure that NVC responses are correct does not imply that these theories view people as inherently irrational, however. Both can still recognize an error when given a counterexample. For example, if VR is given statements describing a counterexample to an erroneous conclusion, it can construct a model of those statements and determine that the premises, but not the conclusion, are consistent with that model.

In terms of aggregate data, mental model theory and VR are both able to account for the major regularities. But VR makes a number of other accurate predictions (Table 6) that have not been derived from mental model theory. In particular, the prediction that time limits should lead to more NVC responses may run counter to model theory (according to model theory, many NVC responses arise only after putative conclusions have been falsified; thus, time limits might be expected to decrease rather than increase the number of NVC responses). Most important, VR accurately models the detailed behavior of individual participants. No other theory, including model theory, has predicted data at such a fine level of detail.

There are other possible ways to distinguish the theories as well. Model theory assumes that participants falsify putative conclusions in the process of deriving the correct answer. If so, participants might be faster or more accurate in rejecting these putative conclusions in comparison with other legal conclusions that they did not explicitly falsify. VR makes the opposite prediction because these possible conclusions are more likely to have been consistent with the annotated model that was constructed.

Finally, the previous falsification analysis showed that falsification was an unnecessary assumption in achieving VR's fits. Only 1 participant's fit (out of 103) improved by more than 5% when VR falsified in comparison with when it did not. The point here is not to show that falsification is unnecessary or irrelevant to deduction. As previously mentioned, this falsification analysis is specific to VR and does not rule out the search for alternative models in general. In particular, mental model theory would not necessarily fit the data better (or equally well) without falsification. Rather, the point is that assumptions about verbal processes (encoding and reencoding) can provide most of the predictive leverage in accounting for categorical syllogism data and that reasoning-specific strategies such as falsification may thus be less important in comparison. A natural question is whether the same is true of other deductive reasoning tasks. We now show that it is.

Verbal Reasoning on Other Deductive Tasks

Aside from verbal reasoning, the mental model theory of syllogistic reasoning (Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991) is the most successful in accounting for the empirical results. But there is another important reason for preferring it to earlier theories: Unlike the others, it has been used to account for behavior on many other reasoning tasks. In fact, mental model theory has now been successfully applied to all of the standard tasks used in studying deductive reasoning (Johnson-Laird & Byrne, 1991). As one would expect, each

individual microtheory makes a few task-specific assumptions, but they are all based on the same general framework.

Mental model theory assumes three major stages in reasoning: *comprehension* (encoding the problem statement into an initial model), *description* (formulating a conclusion based on the model), and *validation* (searching for an alternative model that falsifies the putative conclusion [i.e., falsification]). The major difference between this theory and verbal reasoning is that model theory assumes that falsification is at the heart of deduction, whereas verbal reasoning assumes that reencoding is. VR demonstrates that syllogistic reasoning can be explained without falsification and that verbal processes may thus provide most of the predictive leverage. A natural question involves the importance of falsification in comparison with verbal processes in explaining behavior on other deductive tasks.

We investigate this question by examining the role of falsification in the mental model explanations for other deductive reasoning tasks. We attempt to show that falsification is not critical in explaining the empirical data addressed but that most of the theoretical action is coming from the mental model that is produced by comprehension. Because encoding and reencoding are the processes that lead to this model, such an analysis provides additional support for the verbal reasoning view of reasoning. It demonstrates that verbal reasoning generalizes beyond categorical syllogisms to other deductive reasoning tasks. It also identifies behaviors that require assumptions beyond verbal reasoning and thus provides further insight into the scope of the theory. Following Johnson-Laird and Byrne's (1991) organization, we first consider propositional reasoning. We then focus on conditional reasoning, a specific type of propositional reasoning that has been extensively studied. Next, we discuss relational and quantificational reasoning and, finally, turn to metadeduction.

Propositional Reasoning

Propositional reasoning tasks involve statements that can be assigned a truth value (propositions) and connectives (such as *and, if, or,* and *not*) for composing them into new propositions. Psychologists have mainly focused on two types of propositional reasoning: *conditional reasoning* (involving *if*) and *disjunctive reasoning* (involving *or*). Figure 12 gives typical examples of both types of tasks.

Like categorical syllogisms, both tasks usually consist of two premises. One premise contains a propositional connective (*if p then q, p only if q, p or q*), whereas the other is simply an atomic proposition or its negation (p, q, $-p$, $-q$). In addition to the connective *if*, a conditional premise (*if p then q*) consists of an antecedent (p in *if p then q*) and a consequent (q in *if p then q*). There are four standard inferences that subjects tend to draw on conditional reasoning tasks (aside from "nothing follows"): *modus ponens (MP), modus tollens (MT), denied antecedent (DA),* and *affirmed consequent (AC)* (Figure 12, left). The first two are deductively valid, and the last two are not. Similarly, there are four standard inferences on disjunctive tasks (Figure 12, right), and only two are deductively valid (the first two in the figure). Disjunctions are consistent with both an inclusive and an exclusive interpretation. With an inclusive interpretation, the disjunction is considered true if either (or both) of the constituent propositions is true (*p or q, or both*). An exclusive disjunction is true only if one but not both of the propositions is true (*p or q, but not both*). With an exclusive interpretation, all four of the inferences on the right of Figure 12 are deductively valid. Many studies have added the words *or both* or *but not both* to make the intended interpretation unambiguous.

Now consider the mental model explanation of behavior on these tasks (Johnson-Laird & Byrne, 1991). Again, the goal is to assess the role of falsification in comparison with verbal processes in these explanations. The first findings that were addressed involve the interpretation of disjunctions and conditionals. Neither type of proposition is consistently interpreted by participants. Some experiments find a bias toward interpreting disjunctions as inclusive (Evans & Newstead, 1980), whereas others find a bias toward exclusive interpretations (Manktelow, 1980). Similarly, conditionals are sometimes interpreted as biconditionals (*if and only if p then q*), and sometimes (even by the same participants) they are not. Johnson-Laird and Byrne (1991) explained these findings by assuming that participants construct mental models that do not explicitly represent all of the available information. For a disjunction such as *p or q,* these authors assumed that participants construct two models, one in which p is true and one in which q is true. Such a representation leaves open whether or not both p and q could be true at once (in the same model). Hence, it is consistent with both an inclusive and an exclusive interpretation. For a

Conditional Reasoning		**Disjunctive Reasoning**
If p then q, p only if q		p or q
Modus ponens (MP):	p, therefore q	-p, therefore q
Modus tollens (MT):	-q, therefore -p	-q, therefore p
Denied antecedent (DA):	-p, therefore -q	p, therefore -q
Affirmed consequent (AC):	q, therefore p	q, therefore -p

Figure 12. Propositional reasoning tasks.

conditional such as *if p then q*, Johnson-Laird and Byrne (1991) assumed that participants construct a single explicit model in which p is true and therefore q is true as well. This representation does not make explicit whether there could be other models in which q is true but not p; consequently, it is consistent with both a standard conditional interpretation and a biconditional interpretation. The important point here is that both of these explanations depend critically on the result of verbal processes (i.e., the initial mental model) rather than on falsification. As long as comprehension delivers mental models like those assumed by Johnson-Laird and Byrne (1991), conditionals and disjunctions will be interpreted in an indeterminate way.

It is not particularly surprising that the previous explanations depended more on verbal processes than they did on falsification. After all, they were explanations of how participants interpreted propositions, not how they reasoned with them. The other findings about propositional reasoning that Johnson-Laird and Byrne (1991) addressed, however, involved the difficulty of different propositional inferences. Consider the finding that *modus ponens* is easier than *modus tollens* in the standard conditional reasoning task (Figure 12, left). According to mental model theory, comprehending *if p then q* delivers two models, one in which both p and q are true and one that has no explicit content. Because p is true in the first model, it accommodates the categorical premise for *modus ponens* (p), and the second model (with no explicit content) is eliminated. The remaining first model supports the inference that q is true (the standard *modus ponens* conclusion), and so it should be relatively easy. In contrast, the categorical premise for *modus tollens* ($-q$) leads to the elimination of the first model, leaving only the model with no explicit content (which is updated to reflect that $-q$ is true). Because this model does not support the *modus tollens* conclusion ($-p$), making such an inference should be relatively difficult. But note that the critical assumptions again involve the results of comprehension (whether the resulting models support the conclusion of the inference or not) rather than the search for alternative models (falsification).

The accounts of the other propositional reasoning results similarly depend on assumptions about comprehension rather than falsification. The difference in difficulty of *modus ponens* and *modus tollens* is predicted to disappear when propositions of the form *p only if q* are used, because comprehension is assumed to deliver two explicit models that together support both inferences. And because dealing with two explicit models is assumed to be harder than dealing with one, both inferences are predicted to be harder than *modus ponens* is with the standard conditional, *if p then q*. Similarly, *modus tollens* is predicted to be easier with a biconditional (*if and only if p then q*) than with a standard conditional because the former is assumed to require only two explicit models, whereas the latter is assumed to require three. No such difference is predicted for *modus ponens* because, in both cases, it is assumed to require only one explicit model. Conditionals are predicted to be easier than exclusive disjunctions because they call for the initial construction of only one explicit model, whereas exclusive disjunctions call for the construction of two. Finally, double disjunction tasks such as the following are predicted to be extremely difficult:

> Linda is in Cannes or Mary is in Tripoli, or both.
> Mary is in Havana or Cathy is in Sofia, or both.
> What, if anything, follows?

The reason is that they are assumed to require the initial construction of a very large number of models (five in the preceding task). But if exclusive disjunctions are used instead of inclusive disjunctions (which require an additional model to represent the *or both* situation), the task is predicted to become easier.

The data bear out all of these predictions. But all of the explanations are based on the number of explicit models that are assumed to be delivered by comprehension or on whether or not an initial model supports an inference. The important point here is that verbal processes are providing the theoretical leverage, not the construction of alternative models to falsify a putative conclusion.

Conditional Reasoning

Conditional reasoning has been extensively studied, and many variants of the basic task presented in the last section have been formulated, leading to the discovery of many additional empirical results. Johnson-Laird and Byrne (1991) showed that mental model theory can also account for these other results. Once again, our intent is to assess the role of verbal processes in comparison with falsification in these explanations.

Meaning of conditionals. Mental model theory makes the following assumptions about the way in which conditionals are interpreted (Johnson-Laird & Byrne, 1991, pp. 72–73).

1. An indicative conditional is interpreted by constructing an explicit model of its antecedent, which is exhaustively represented and to which is added a model of the consequent. An alternative implicit model allows for cases in which the antecedent does not hold.

2. A counterfactual conditional is interpreted in the same way, except that the models of its antecedent and consequent are of counterfactual situations and there is an explicit model of the actual situation.

3. Conditionals may elicit richer models in which more states are rendered explicit. This fleshing out of models occurs in several circumstances (e.g., when a referential relation, or one based on general knowledge, holds between the antecedent and consequent).

A conditional is considered true if the consequent is true in the context asserted by the antecedent and false if the consequent is false in such a context. This theory is consistent with a number of aspects of the everyday semantics of conditionals. For example, when untrained participants negate a conditional, they usually negate the consequent and leave the antecedent unchanged (*if p then q* becomes *if p then not q*), presumably because the context (defined by the antecedent) is assumed to remain constant. This also explains why people often judge conditionals to be irrelevant in situations in which the antecedent is false (although, logically, such conditionals are true) and why they assert conditionals only if there is some reason for relating the antecedent and consequent (e.g., people do not say, "If grass is green, then the earth is round," even though it is logically true).

Johnson-Laird and Byrne (1991) went into much greater de-

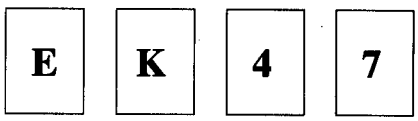

Figure 13. Wason's selection task.

tail, describing the specific models they believed are built for each type of conditional, making explicit what they meant by all of the terms in the summary just discussed (e.g., *exhaustively represented*), and showing how this theory accounts for other aspects of the way in which people understand and use conditionals. Even without going into the details, however, it is clear that the theory depends on how conditionals are interpreted during comprehension rather than on falsification. All three points in the preceding summary relate to verbal processes or their results; they do not make reference to the attempt to falsify conclusions by searching for alternative models.

Of course, it is only natural that the mental model theory for the meaning of conditionals should be based on assumptions about comprehension rather than on assumptions about falsification. Johnson-Laird and Byrne (1991) went on to address how people reason with conditionals—the paradoxes of implication, the empirical results surrounding Wason's (1966) selection task, and the suppression of valid conditional inferences—and we now turn to these results.

Paradoxes of implication. First, consider what Johnson-Laird and Byrne (1991) called the paradoxes of implication. As long as its antecedent is false, a conditional is guaranteed to be valid, regardless of its consequent. For example, "If grass is red, then Bush is president" is a logically valid proposition (because grass is not red). Similarly, as long as its consequent is true, a conditional with any antecedent whatsoever is guaranteed to be valid (e.g., "If I am president, then grass is green"). Intuitively, however, neither proposition seems valid, even though logically both are valid. Johnson-Laird and Byrne (1991) argued that the reason is that these propositions throw away semantic information. That is, there are more states of affairs in which the conditional is true than in which the negated antecedent (in the first case) or consequent (in the second) is true. Hence, "Grass is not red" is more constraining than "If grass is red, then Bush is president," and "Grass is green" is more constraining than "If I am president, then grass is green." People would never use the conditional in such cases. Here again, the search for alternative models does not play a role in accounting for these paradoxes.

Wason's selection task. In Wason's (1966) selection task, participants are shown four cards and are told that each card has a letter on one side and a number on the other (Figure 13). The upper (visible) faces of the cards contain the symbols *E, K, 4,* and *7.* The participants are then presented with a conditional proposition such as the following:

> If a card has a vowel on one side,
> then it has an even number on the other.

Their task is to indicate those cards (and only those cards) that must be turned over to determine whether the rule is true or false.

One of the reasons that Wason's (1966) selection task has attracted so much attention is that it is so difficult. The correct behavior is to select the *E* and *7* cards, and yet the vast majority of participants either select only the *E* card or select the *E* and *4* cards. Performance can be facilitated by using certain types of realistic content. For example, if a conditional such as the following is used:

> If someone is drinking beer then they must be over 18

and if one side of the cards specifies the age of a person in a bar and the other side specifies whether that person is drinking beer or a nonalcoholic drink, participants perform better (Griggs & Cox, 1982).

Johnson-Laird and Byrne (1991) summarized the model theory of the selection task as follows:

1. Participants consider only those cards that are explicitly represented in their models of the rule.
2. They then select those cards for which the hidden value could have a bearing on the truth or falsity of the rule.

The authors went on to derive all of the major regularities in the selection task from these simple assumptions.

Once again, note that neither assumption makes reference to the search for alternative models. So does the mental model theory constitute a verbal reasoning account of behavior on the selection task? The explanations of the findings that we have previously discussed were all based solely on assumptions about comprehension and description and could thus be clearly characterized as verbal reasoning. The same cannot be said for Wason's (1966) selection task. Neither of the assumptions just described involves how participants comprehend the problem statement or how they describe their mental model (although such assumptions also play a major role in their explanations); rather, they involve what hypotheses (about the backs of cards) participants will entertain and how participants will test those hypotheses. As Evans (1982) put it, Wason's selection task is not simply a deductive reasoning problem but a metainference task: "Subjects are not simply required to draw or assess immediate inferences. Rather, they are invited to entertain alternative hypotheses with respect to the truth or falsity of a rule, and asked to test these hypotheses" (p. 157). Johnson-Laird and Byrne (1991) made a similar point: "The selection task calls for more than a deduction: subjects have to explore different possibilities, to deduce their consequences for the truth or falsity of the rule, and on this basis to determine which cards to select" (p. 75).

It should not be surprising, then, that the selection task requires more than verbal reasoning. The reason is that the task requires knowledge or strategies that are not explicitly provided in the problem statement. As a result, the processes of comprehension and description, although important, are not sufficient. More generally, verbal reasoning will not generalize to any task

that requires knowledge that is not available in the problem statement (specifically in a verbal format), and this provides an important characterization of the scope of verbal reasoning as a theory. Interestingly, whereas purely deductive tasks almost always satisfy this criterion, some nondeductive tasks do as well. As a trivial example, consider a modified categorical syllogism in which the task is to generate a conclusion that is possible (rather than deductively valid) given the premises. Such a task does not require any knowledge outside of the problem statement, and yet it is clearly not a deductive reasoning task. A theory based on verbal reasoning (e.g., a version of VR modified to reflect the different demands of the task) could apply in such situations.

Suppression of valid deductions. Byrne (1986, 1989) has found that, under certain circumstances, even the most natural inferences (e.g., *modus ponens*) can be suppressed. She used sets of premises such as the following:

> If she meets her friends then she will go to a play.
> If she has enough money then she will go to a play.

When participants are presented with such premises and only one of the antecedents ("she meets her friends"), they tend not to make the valid *modus ponens* inference ("she will go to a play"). Johnson-Laird and Byrne (1991) argued that, under these conditions, comprehension delivers one explicit model in which both antecedents and the consequent are true and one model with no explicit content. Because only one of the antecedents is asserted, the first model is rejected in favor of the second, and no *modus ponens* inference is made. As in many of the previous cases, this explanation is derived from assumptions about the models that comprehension delivers; it does not depend on falsification.

Spontaneous use of conditional descriptions. Johnson-Laird and Byrne (1991) also addressed how conditionals are used in descriptions. Byrne and Johnson-Laird (1990, 1992) presented participants with factual sentences such as the following:

> Laura has an essay to write.
> The library stays open.
> Laura studies late in the library.

When asked to combine them into a single sentence, participants used conditionals on only 2% of trials; they tended to use connectives such as *and* or *when* instead. Johnson-Laird and Byrne assumed that factual statements such as those just mentioned lead to a single explicit model in which all of the assertions are true. According to the model theory, the representation of conditionals includes a model with no explicit content, and so conditionals would not be used to summarize such statements. However, model assertions such as "Laura *can* study late in the library" are, in fact, predicted to lead to the creation of such implicit models, and so conditionals should be used more frequently. Consistent with this prediction, conditionals appeared in 36% of such trials (as compared with 2% on the other trials). In this case, the critical assumptions involve description: How will different types of models be described? (Specifically, when will conditionals be used in describing them?) Thus, this explanation is another example of a verbal reasoning account, and falsification does not play a central role.

Problem I	**Problem II**
A is on the right of B C is on the left of B D is in front of C E is in front of B What is the relation between D and E?	B is on the right of A C is on the left of B D is in front of C E is in front of B What is the relation between D and E?
Problem III	**Problem IV**
B is on the right of A C is on the left of B D is in front of C E is in front of A What is the relation between D and E?	A is on the right of B C is on the left of B D is in front of C E is in front of A What is the relation between D and E?

Figure 14. Two-dimensional spatial reasoning tasks.

Reasoning About Relations

Three-term series problems (also termed linear syllogisms) involve reasoning about relations. The following is a typical example.

> John is taller than Bill.
> Mary is shorter than Bill.
> Therefore, John is taller than Mary.

A major issue in the study of such tasks has been whether behavior can best be explained in terms of linguistic factors (Clark, 1969) or by assuming the construction of a spatial array (Huttenlocher, 1968). Johnson-Laird and Byrne (1991) considered Clark's linguistic account to be an example of a rule theory (that uses context-specific rules), and they considered Huttenlocher's imagery account to be a model theory. They argued that these tasks do not have enough structure to distinguish these different accounts, and so they turned their attention to two-dimensional spatial deductions such as those shown in Figure 14. They showed that, for such tasks, theories based on rules and theories based on models make different predictions and that empirical results support model theory rather than rule theory.

For example, Hagert's rule theory (1983, 1984) predicts that Problems I and II in the figure should be equally difficult because the formal derivations of the conclusions contain the same number of steps. In contrast, model theory predicts that Problem I should be easier because it is consistent with only one model of the premises:

C B A
D E

whereas Problem II is consistent with two:

C A B A C B
D E D E

Although this prediction depends on the number of models, it nevertheless does not rely on falsification. The reason is that the two models for Problem II both support the same conclusion about *D* and *E*: that *D* is to the left of *E*. The second model does not falsify the initial conclusion; rather, it confirms it. Clearly,

then, this prediction is based not on searching for an alternative model that falsifies the putative conclusion but on comprehension delivering ambiguous results.

One prediction in which falsification does play a central role, however, is that Problem III in Figure 14 should be harder than both Problems I and II; this problem requires falsification, and the other two problems do not. An initial model of Problem III such as

C A B
D E

leads to the conclusion "D is left of E." But this conclusion can be falsified by considering the alternative model:

A C B
E D

In keeping with this prediction, only 18% of participants solve tasks like Problem III correctly (as compared with 61% and 50% for Problems I and II). According to model theory, the reason participants solve this task incorrectly is that they fail to successfully search for an alternative model. That is, according to the theory itself, the vast majority of the time participants do not successfully apply the falsification stage to these problems. And even the few participants who correctly solve such tasks may not be falsifying. If they noticed the ambiguity of the premises during encoding (as the authors assumed they do for Problem II), then they might very well respond "NVC" without having constructed alternative models.

Finally, mental model theory predicts that Problem IV in Figure 14 should be relatively easy because it supports only one model:

C B A
D E

This prediction is indeed borne out in the data (70% of participants solve this task correctly). But, again, the explanation relies on the results of comprehension (a verbal process) rather than on falsification.

In summary, assumptions about linguistic processes and their results, not falsification, provide the theoretical leverage in almost every explanation. And in the explanation in which falsification does play the central role, model theory itself assumes that participants rarely (less than 20% of the time) succeed in applying the strategy. Finally, even the behavior of those participants that model theory assumes are falsifying can be explained by an alternative verbal reasoning account.

Categorical Syllogisms

Johnson-Laird and Byrne (1991) discussed two issues that VR has not addressed, and so it is worth considering whether either of them provides strong evidence that falsification, rather than verbal processes, provides most of the theoretical leverage. The first involves participants' memory of their responses, and the second involves the effects of using *only* as the quantifier.

Byrne and Johnson-Laird (1989) had participants solve a set of 16 syllogisms and subsequently asked them to choose the response they had given from a set of four alternatives. The alternatives never included "NVC" as a choice, even though this was the correct answer on half of the problems and many participants had, in fact, used it as their response. On 74% of the trials in which participants had correctly responded "NVC," they chose the response that was consistent with considering only a single model. The authors suggested that this behavior can best be explained by assuming that the participants initially considered that response but rejected it when they constructed a falsifying model; they viewed it as evidence that falsification is playing a central role in deduction. This behavior can be explained in terms of verbal processes, however, by assuming that the participants were simply solving these problems again. Johnson-Laird and Byrne (1991) acknowledged as much: "[The possibility] that they were reasoning from the premises once again . . . cannot be eliminated" (p. 127).

Such a strategy is obviously plausible for participants who did not remember their previous response. But it even makes sense for those who did because they would not find it among the alternatives listed, so solving the problem again would be a natural way to proceed. Furthermore, participants who did remember their previous response might also be expected to remember having falsified the conclusion based on the initial model. If so, they might be less likely rather than more likely to choose that response, in opposition to the empirical finding.

The second issue involves the quantifier *only*. The premise *Only the a's are b's* is logically equivalent to *All the b's are a's*, but Johnson-Laird and Byrne (1989) assumed that *Only the a's are b's* leads to the explicit representation of negative information (specifically that anything that is *not* an *a* is also *not* a *b*), whereas *All the b's are a's* does not. They then argued that a model theory augmented with this assumption can explain a variety of empirical results.

The first empirical finding they addressed is that problems using *only* are reliably harder than those using *all* (26% vs. 46% correct in Johnson-Laird & Byrne, 1989). They argued that the reason is that "the [initial] model for 'only' is more complex than the one for 'all' " (Johnson-Laird & Byrne, 1991, p. 129); a verbal reasoning account based on the results of comprehension rather than on falsification. Second, problems that did not require falsification were much easier (55% correct) than those that did (and had a valid conclusion; 15% correct). As in the case of relational reasoning, the model theory itself assumes that participants successfully falsify their conclusions less than 20% of the time. Also, because valid conclusions are guaranteed to be true in all models of the premises, they are guaranteed to be true in any initial model the participants may construct. Thus, the participants who did solve these problems correctly may simply have generated the valid conclusion from the initial model without ever having constructed alternative models. Third, when both premises contained the quantifier *only*, just 16% of conclusions used it, a result that runs counter to the standard atmosphere effect in syllogistic reasoning. Johnson-Laird and Byrne (1989) suggested that, because these models are also consistent with *all* conclusions, participants prefer using that quantifier because it does not require processing negative information. Again, these are assumptions about comprehension and description rather than falsification, and so this too is a verbal reasoning explanation. Finally, when Johnson-Laird and Byrne (1989) presented participants with premises such as

A one-model problem

None of the painters is in the same place as any of the musicians.
All of the musicians are in the same place as all of the authors.
Therefore, none of the painters is in the same place as any of the authors.

A multiple-model problem

None of the painters is in the same place as any of the musicians.
All of the musicians are in the same place as some of the authors.
Therefore, none of the painters is in the same place as some of the authors.

Figure 15. Multiply quantified deductive reasoning tasks.

All authors are bankers.
Mark is an author.

and

Only bankers are authors.
Mark is not a banker.

they found an interaction between the quantifier (*all* or *only*) and the polarity or quality of the second premise (whether or not it was negated). Specifically, when *all* was used, participants performed significantly worse if the second premise was negated than if it was not (73% vs. 96% correct). But if *only* was used, no such difference was found (86% vs. 90% correct). According to mental model theory, the reason is that *only,* but not *all,* leads to the explicit representation of negative information (about nonbankers). Consequently, inferring that Mark is not an author (in the second example) should be just as easy as inferring that an author is a banker; both are supported in the initial model. In contrast, the initial model for *all* supports the inference only from an affirmative premise, and so making inferences from negated premises should be harder. Again, however, the critical assumptions are about the initial model delivered by comprehension, not about falsification.

Reasoning With Multiple Quantifiers

Next consider multiply quantified deductions such as those shown in Figure 15. Johnson-Laird, Byrne, and Tabossi (1989) presented results from three experiments designed to distinguish mental model theory from theories based on formal rules. Their data run counter to rule theories but are consistent with model theory.

The main result in all three experiments was the same: Problems that are consistent with multiple models are harder than those that are consistent with only one. In this case, falsification is playing the central explanatory role, rather than verbal processes; to solve the multiple model problems correctly, participants must consider alternative models that falsify their initial conclusions. Because there are no such alternative models to consider in one-model problems, these tasks place less of a load on working memory and are consequently easier.

Can these results be explained in terms of verbal reasoning rather than falsification? Consideration of a number of points illustrates that they can. First of all, as in previous cases, very few participants correctly solved the problems that were assumed to require falsification. On valid multiple-model problems in the three experiments (those that required falsification), only 13%, 16%, and 23% of responses were correct. Thus, once again, according to the model theory itself, participants do not successfully falsify their conclusions very often. Participants performed fairly well on invalid multiple-model problems (those with no valid conclusion), correctly solving 50%, 40%, and 23% of these tasks in the three experiments. But this result cannot be attributed to the use of falsification because that would not predict the large difference in performance between valid and invalid multiple-model problems (if participants can successfully falsify on the invalid problems, they should also be able to falsify on the valid problems).

Verbal reasoning provides a natural account of the difference between one-model and multiple-model problems. The reason one-model problems are easy is that any initial model of the premises will support only valid conclusions (otherwise, they could be falsified by an alternative model, and it would not be a one-model problem). In contrast, the initial models for multiple-model problems can support invalid conclusions (that is why they can be falsified by alternative models), and so they should be harder. Furthermore, even the rare cases that seem to suggest falsification (correct responses to valid multiple-model problems) can be explained more simply without it. Rather than constructing a sequence of alternative models and finding a conclusion that is true in all of them, participants may simply be generating the valid conclusion from the initial model. After all, because the conclusion is valid, it is guaranteed to be supported in any model of the premises, including the one that is constructed initially. Similarly, correct NVC responses to invalid syllogisms do not necessarily imply a search for alternative models; participants either may be unable to construct a model that supports any legal conclusion or may notice some ambiguity during encoding and decide to respond NVC rather than risk a conclusion that they realize may not be valid. In short, all of these results can be explained naturally and parsimoniously in terms of verbal reasoning.

More generally, the preceding argument shows that, for any task, a difference in difficulty that is predicted on the basis of the use of falsification (one vs. many models) can be explained more simply by a verbal reasoning account. Participants may find multiple-model problems more difficult, not because they search through a sequence of alternative models but because the initial model they construct supports invalid conclusions. By definition, the initial model for one-model problems cannot support invalid conclusions, and so these problems are easier.

Metadeduction

Metalogical reasoning problems make explicit reference to truth and falsity. For example, consider the following "knight-and-knave" puzzle:

> Knights always tell the truth while knaves always lie. Lancelot says, "I am a knave and so is Gawain." Gawain says, "Lancelot is a knave." What are Lancelot and Gawain?

The correct answer is that Lancelot is a knave (he is lying when

he says Gawain is a knave, and hence the conjunction is also a lie) and Gawain is a knight.

Johnson-Laird and Byrne (1991) proposed that participants use certain metalogical strategies in addition to their standard processes for constructing a model, generating a conclusion from a model, and searching for alternative models. Johnson-Laird and Byrne summarized these strategies as follows:

1. Simple chain: Assume that the assertor in the first premise tells the truth and follow up the consequences, but abandon the procedure if it becomes necessary to follow up disjunctive consequences. Assume that the assertor in the first premise is lying and do likewise.

2. Circular: If a premise is circular, follow up the immediate consequences of assuming that it is true and then follow up the immediate consequences of assuming that it is false.

3. Hypothesize and match: If the assumption that the first assertor A is telling the truth leads to a contradiction, then attempt to match $-A$ with the content of other assertions, and so on.

4. Same assertion and match: If two assertions make the same claim and a third assertor, C, assigns the two assertors to different types, or vice versa, then attempt to match $-C$ with the content of other assertions, and so on.

On the basis of these strategies and the basic model theory, Johnson-Laird and Byrne (1991) went on to derive three predictions about metalogical reasoning. For our purposes, the specific details of the theory (e.g., exactly how the preceding strategies work) are not important. What is important is the role that falsification plays in the explanations.

The first prediction is that problems that can be solved using one of the preceding four strategies will be easier than problems that cannot. Consistent with this prediction, problems for which one of the strategies was sufficient were correctly solved 28% of the time, in comparison with only 14% correct responses on other problems. This prediction is based on the specific metalogical strategies just discussed, none of which make reference to falsification. Johnson-Laird and Byrne (1991) stated the second prediction as follows: "The difficulty of a problem will depend on the number of clauses that it is necessary to use in order to solve the problem" (p. 160). They assumed that additional clauses put extra strain on working memory and make the problem more difficult. They went on to spell out what they meant by using a clause (basically making one inferential step); regardless of the details, however, it is clear that this prediction does not depend on searching for alternative models. Indeed, they acknowledged that it really does not depend on the basic model theory at all: "The prediction is almost independent of the processing theory that we have proposed, and is likely to be made by any sensible analysis of meta-logical problems" (p. 160). The third prediction is that the hypothesis that an assertion is true should be easier to process than the hypothesis that an assertion is false. The reason is that negation causes problems. Again, this prediction does not depend on falsification.

A second kind of metadeduction involves deducing what someone else could have deduced. For example, consider the following problem:

> Three wise men who were perfect logicians were arrested by the Emperor on suspicion of subversion. He put them to the following test. The three men were lined up in a queue facing in the same direction, and a hat was placed on the head of each of them. The men could not see their own hats, but the man at the back of the queue (A) could see the two hats in front of him, the man in the middle (B) could see the one hat in front of him, and the man at the front (C) could see no hat. The Emperor said: "If one of you can tell me the colour of your own hat, I will set all three of you free. There are three white hats and two black ones from which your hats have been drawn. I will now ask each of you if he can tell me the colour of his hat. You may answer only 'yes', 'no', or 'I don't know'." A who could see the two hats in front of him said, "I don't know." B heard A's answer and said, "I don't know." C heard the two previous answers. What was C's answer? (Johnson-Laird & Byrne, 1991, p. 162).

It turns out that C can deduce the color of his hat (it must be white) on the basis of the answers of A and B (who are perfect logicians). A could not have seen two black hats; otherwise, he could have deduced that his own hat was white. Consequently, one or both of B and C are wearing a white hat. If C's hat were black, then B could have deduced that his own hat was white. Since he did not, C knows that his hat must be white.

The only empirical result about tasks like this that mental model theory addresses is the fact that they are difficult. Johnson-Laird and Byrne (1991) proposed that there are two major sources of difficulty: working memory load and the lack of an appropriate strategy. The preceding task requires constructing models of C's models, which are, in turn, models of A's and B's models of the situation. Keeping track of all of this information undoubtedly strains the limits of working memory. And the appropriate strategy for solving the problem is far from obvious at first glance. For our purposes, however, the question is whether this explanation of the difficulty of these tasks depends on falsification. And neither the amount of information that must be represented nor the lack of an appropriate strategy involves the search for alternative models.

This fact does not imply that a verbal reasoning account will be sufficient to explain these phenomena, however. As we emphasized in the discussion of Wason's (1966) selection task, verbal reasoning will not generalize to tasks that require knowledge that is not available in the problem statement in a verbal format. And the knowledge incorporated in the strategies proposed earlier is not. As a result, the processes of comprehension and description are not sufficient to account for behavior on this task.

Discussion

We have now reviewed all six of the standard deductive reasoning domains: propositional reasoning, conditional reasoning, relational reasoning, syllogistic reasoning, reasoning with multiple quantifiers, and metadeduction. In some cases (propositional reasoning and most of conditional reasoning), we showed that the model theory's explanations depended on comprehension and description, rather than falsification, and could thus be reinterpreted as verbal reasoning accounts. In other cases (syllogistic reasoning and reasoning with multiple quantifiers), we argued that model theory itself assumed that falsification was rare and showed that alternative accounts based on verbal reasoning could account for the behavior. In

short, these analyses demonstrated that verbal reasoning can account for behavior across a wide variety of reasoning tasks; it has extensive breadth.

But these analyses also showed that verbal reasoning cannot account for behavior that depends on knowledge not provided verbally in the problem statement. Johnson-Laird and Byrne's (1991) explanations of behavior on Wason's (1966) selection task and in metadeduction depended on assumptions beyond both model theory and verbal reasoning (specifically, about the types of strategies that participants use), probably because both tasks involve metainference. Linguistic processes are not sufficient to account for such behavior. On all of the standard deductive reasoning tasks (which do not involve metainference), however, verbal reasoning provides accurate accounts of human behavior.

Conclusion

Previous theories of deduction have shared the same basic structure at the most general level: Participants encode the problem statement into some internal representation, apply certain reasoning-specific processes to that representation (e.g., searching for alternative models or applying formal or content-specific rules of inference), and then decode the result. We propose that the central processes in deductive reasoning are linguistic (encoding, reencoding, and generation) rather than reasoning-specific skills.

This verbal reasoning hypothesis provides a parsimonious and accurate account of deductive reasoning. It explains behavior in terms of standard linguistic processes without the need to posit reasoning-specific mechanisms. When implemented in the form of a computational model of syllogistic reasoning, it provides the most detailed and accurate account to date of behavior on the task; it explains all of the major phenomena, it makes accurate novel predictions, and it models the behavior of individual participants with an accuracy that rivals their own test–retest reliability.

Verbal reasoning also explains behavior on a variety of other deductive reasoning tasks. Mental model theory accounts of many of these tasks depend on the results of comprehension and description, rather than falsification, and thus constitute verbal reasoning explanations. In the cases in which falsification does play an explanatory role, simpler accounts based on purely linguistic processes also explain the behavior. Verbal reasoning thus extends to almost all deductive tasks. The theory is not sufficient to explain behavior on reasoning tasks that rely on information that is not provided verbally in the problem statement, however (e.g., metainference and many types of nondeductive reasoning); such behavior must be explained with reference to nonlinguistic processes. But on tasks in which all of the relevant information is so provided (as is the case on almost all deductive tasks), verbal reasoning provides a compelling new view of behavior.

References

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Begg, I., & Denny, P. (1969). Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. *Journal of Experimental Psychology, 81,* 351–354.

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

Byrne, R. M. J. (1986). *The contextual nature of conditional reasoning.* Unpublished doctoral dissertation, University of Dublin.

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

Byrne, R. M. J., & Johnson-Laird, P. N. (1989). *Re-constructing inferences.* Unpublished manuscript, University of Wales College of Cardiff.

Byrne, R. M. J., & Johnson-Laird, P. N. (1990). Models and deduction. In K. Gilhooly, M. T. G. Keane, R. Logic, & G. Erdos (Eds.), *Lines of thought: Reflections on the psychology of thinking* (Vol. 1, pp. 139–151). New York: Wiley.

Byrne, R. M. J., & Johnson-Laird, P. N. (1992). *The spontaneous use of propositional connectives. Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 45A,* 89–110.

Ceraso, J., & Provitera, A. (1971). Sources of error in syllogistic reasoning. *Cognitive Psychology, 2,* 400–410.

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

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

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review, 76,* 387–404.

Dickstein, L. S. (1975). Effects of instruction and premise order on errors in syllogistic reasoning. *Journal of Experimental Psychology: Human Learning and Memory, 104,* 376–384.

Erickson, J. R. (1974). A set analysis theory of behavior in formal syllogistic reasoning tasks. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 305–329). Potomac, MD: Erlbaum.

Evans, J. S. B. T. (1982). *The psychology of deductive reasoning.* London: Routledge & Kegan Paul.

Evans, J. S. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11,* 295–306.

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

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

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

Gonzalez-Marques, J. (1985). La influencia de materiales no emocionales en la solucion de silogismos categoricos [The effects of nonemotional stimulus materials on the solution of categorical syllogisms]. *Informes-de-Psicologia, 4,* 183–198.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic materials effect in the Wason 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.

Hagert, G. (1983). *Report of the Uppsala programming methodology and artificial intelligence laboratory.*

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

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

Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. *Psychological Review, 75,* 550–560.

Inder, R. (1986). Modeling syllogistic reasoning using simple mental models. In A. G. Cohn & J. R. Thomas (Eds.), *Artificial intelligence and its applications* (pp. 211–225). New York: Wiley.

Inder, R. (1987). *The computer simulation of syllogism solving using*

restricted mental models. Unpublished doctoral dissertation, University of Edinburgh, Edinburgh, Scotland.

Janis, I. L., & Frick, F. (1943). The relationship between attitudes towards conclusions and errors in judging logical validity of syllogisms. *Journal of Experimental Psychology, 33,* 73–77.

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

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *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.

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

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence, 33,* 1–64.

Lee, J. R. (1983). *Johnson-Laird's mental models: Two problems.* Unpublished manuscript, School of Epistemics, University of Edinburgh.

Lehman, J. F., Newell, A., Polk, T., & Lewis, R. L. (1993). The role of language in cognition: A computational inquiry. In G. Harman (Ed.), *Conceptions of the human mind* (pp. 39–58) Hillsdale, NJ: Erlbaum.

Manktelow, K. I. (1980). *The role of content in reasoning.* Unpublished doctoral dissertation, Plymouth Polytechnic University, Plymouth, Devon, England.

Morgan, J. I. B., & Morton, J. T. (1944). The distortions of syllogistic reasoning produced by personal connections. *Journal of Social Psychology, 20,* 39–59.

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

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

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

Polk, T. A. (1992). *Verbal reasoning.* Unpublished doctoral dissertation, Carnegie-Mellon University, Pittsburgh, PA.

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

Polk, T. A., & Newell, A. (1988). Modeling human syllogistic reasoning in Soar. In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 181–187). Hillsdale, NJ: Erlbaum.

Polk, T. A., Newell, A., & Lewis, R. L. (1989). Toward a unified theory of immediate reasoning in Soar. In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 506–513). Hillsdale, NJ: Erlbaum.

Polk, T. A., Newell, A., & VanLehn, K. (1995). *Analysis of symbolic parameter models (ASPM): A new model-fitting technique for the cognitive sciences.* Manuscript in preparation.

Polk, T. A., VanLehn, K., & Kalp, D. (1995). ASPM2: Progress toward the analysis of symbolic parameter models. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 127–139). Hillsdale, NJ: Erlbaum.

Revlis, R. (1975a). Syllogistic reasoning: Logical decisions from a complex data base. In R. J. Falmagne (Ed.), *Reasoning: Representation and process* (pp. 93–133). Hillsdale, NJ: Erlbaum.

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

Revlis, R., Ammerman, K., Petersen, K., & Leirer, V. (1978). Category relations and syllogistic reasoning. *Journal of Educational Psychology, 70,* 613–625.

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

Sells, S. B. (1936). The atmosphere effect: An experimental study of reasoning. *Archives of Psychology, No. 200.*

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116,* 250–264.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135–151). New York: Penguin Books.

Wilkins, M. C. (1928). The effect of changed material on the ability to do formal syllogistic reasoning. *Archives of Psychology, No. 102.*

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

Appendix A

Indirect Knowledge Extracted by VR3

Target Proposition	Reference Property: Y	-X	-Y
All X are Y	All Y are X	No non-X are Y	No non-Y are X
Some X are Y	Some Y are X	Some non-X are Y	Some non-Y are X
No X are Y	No Y are X	No non-X are Y	All non-Y are X
Some X are not Y	Some Y are not X	Some non-X are not Y	Some non-Y are not X

Figure A1. Indirect knowledge extracted by VR3.

Appendix B

Predicted Responses From VR Defaults (VR1, VR2, and VR3)

Syllogism	VR1	VR2	VR3	Syllogism	VR1	VR2	VR3	Syllogism	VR1	VR2	VR3	Syllogism	VR1	VR2	VR3
AabAbc	Aac	Aac	Aac	AabAcb	NVC	NVC	Iac,Aca	AbaAbc	NVC	NVC	Aca	AbaAcb	Aca	Aca	Aca
AabIbc	Iac	Iac	Iac	AabIcb	NVC	NVC	Iac,Ica	AbaIbc	NVC	Ica	Ica	AbaIcb	Ica	Ica	Ica
AabEbc	Eac	Eac	Eac	AabEcb	NVC	Eac	Eca	AbaEbc	NVC	NVC	Eac	AbaEcb	NVC	NVC	Eca
AabObc	Oac	Oac	Oac	AabOcb	NVC	NVC	Oca	AbaObc	NVC	NVC	Oac	AbaOcb	NVC	NVC	Oca
IabAbc	Iac	Iac	Iac	IabAcb	NVC	NVC	Iac,Ica	IbaAbc	NVC	Iac	Ica	IbaAcb	NVC	NVC	Ica
IabIbc	Iac	Iac	Iac	IabIcb	NVC	NVC	NVC	IbaIbc	NVC	Ica	Ica	IbaIcb	NVC	NVC	NVC
IabEbc	Oac	Oac	Oac	IabEcb	NVC	Oac	Iac,Ica	IbaEbc	NVC	Oac	Oac	IbaEcb	NVC	Oac	Ica
IabObc	Oac	Iac	Oac	IabOcb	NVC	NVC	Iac,Ica	IbaObc	NVC	Oac	Oac	IbaOcb	NVC	NVC	Ica
EabAbc	NVC	NVC	Eca	EabAcb	NVC	Eca	Eca	EbaAbc	NVC	NVC	Eca	EbaAcb	Eca	Eca	Eca
EabIbc	NVC	Oca	Oca	EabIcb	NVC	Oca	Oca	EbaIbc	NVC	Oca	Oca	EbaIcb	Oca	Oca	Oca
EabEbc	NVC	NVC	Eac	EabEcb	NVC	NVC	Iac,Aca	EbaEbc	NVC	NVC	NVC	EbaEcb	NVC	NVC	Eca
EabObc	NVC	NVC	Oac	EabOcb	NVC	NVC	Iac,Ica	EbaObc	NVC	NVC	NVC	EbaOcb	NVC	NVC	Oca
OabAbc	NVC	NVC	Oca	OabAcb	NVC	NVC	Oca	ObaAbc	NVC	NVC	Oca	ObaAcb	NVC	NVC	Oca
OabIbc	NVC	NVC	Oca	OabIcb	NVC	NVC	Oca	ObaIbc	NVC	Oca	Oca	ObaIcb	NVC	NVC	NVC
OabEbc	NVC	NVC	Oac	OabEcb	NVC	NVC	Oca	ObaEbc	NVC	NVC	NVC	ObaEcb	NVC	NVC	Oca
OabObc	NVC	NVC	NVC	OabOcb	NVC	NVC	Oca	ObaObc	NVC	NVC	NVC	ObaOcb	NVC	NVC	Oca

Note. A = all; I = some; E = no; O = some not; NVC = no valid conclusion. VR = verbal reasoning model.

Appendix C
Parameters in VR

We assume that the premises *Some x are y* and *Some x are not y* are often interpreted to mean more than simply "there exists an object that has properties X and $Y(-Y)$. Parameters 1 and 2 (see Figure 7) control what additional information is extracted. If they are set to Value a, then VR also ensures that there are other objects with property X that do not have Y or $-Y$ (they may or may not be Y). This interpretation corresponds to the default encoding. With Value b, VR creates a new (different) object that has only property X (it may or may not be Y). Value c is similar except that the new object also has property $-Y$. Value d leads to two other objects with property X, one with $-Y$ and one that may or may not be Y. Finally, with Value e, no additional information is extracted. We assume that the compound semantics of the universal premises (*All* and *No*) are much less ambiguous, and so we do not have such parameters for these premises.

The distinction between different and other objects is subtle but important. It basically corresponds to whether the additional information is assumed to have the same referent as the primary information (e.g., whether or not *Some x may or may not be y* refers to the same set of objects with property X as does *Some x are y*). If they have the same referent (other rather than different), then any other properties that the

(*Appendix C continues on next page*)

referent has (e.g., Z) will appear in both objects. If they do not have the same referent (different), then the additional information will create a completely new object with at most two properties (X and Y or $-Y$). There may be cases in which the primary referent actually contradicts the additional information [e.g., the referent (X Y) contradicts *Some x are not y*]. Under these circumstances, even interpretations corresponding to "other" will create new objects.

The next four parameters (3–6) specify the atomic semantics of the four premise types independent of any additional information that may be extracted. For the universal premises (Parameters 3 and 5), both interpretations augment all existing Xs with property Y (or $-Y$), with Value b, however, they also create a new object (this additional object encodes the fact that there may be additional Xs that have not yet appeared in the annotated model). For the particular premises (Parameters 4 and 6), the alternative interpretations differ in what Xs they refer to (i.e., in what model object ends up becoming augmented). Consider Parameter 4. With Value a, the most recently accessed (MR) object with properties X and Y is chosen, and X is marked as identifying. If no such object exists, then the most recently accessed object with X but not with $-Y$ is augmented. If, again, there is no such object, then a new object (X' Y) is created. This interpretation corresponds to the default. Interpretation b is similar, but it immediately looks for an X that lacks $-Y$, and c gives up (and creates a new object) if it cannot find an object with both X and Y. Value d always creates a new object.

Parameters 7 and 8 control the generation templates for *Some x are y* and *Some x are not y*. Value a for Parameter 7 corresponds to *Some but not (necessarily) all x are y;* in addition to an X that is a Y, there must be an X that is not known to be a Y (it could be a $-Y$ but does not have to be). Value b corresponds to *Some and possibly all;* even if all Xs are Ys, the particular conclusion will be proposed. Parameter 9 controls whether generation can produce conclusions about secondary (nonidentifying) properties. With Value a, generation never produces conclusions about secondary properties. With Value b, secondary properties will be tried, but only after identifying properties.

Parameters 10 through 21 control what indirect knowledge is extracted during reencoded. The first four (10–13) specify what is extracted when the reference property is y and the premise being reencoded relates X to Y. The other two sets of four reencoding parameters specify what is extracted about $-X$ (14–17) and $-Y$ (18–21).

Finally, Parameter 22 controls falsification. With Value a, VR does not try to falsify its putative conclusions. With Value b, it does try to falsify, and, if it succeeds, it responds with NVC.

Received April 7, 1994
Revision received November 30, 1994
Accepted February 9, 1995 ■