Journal of Experimental Psychology: Applied
2001, Vol. 7, No. 4, 308–316

Copyright 2001 by the American Psychological Association, Inc.
1076-898X/01/$5.00 DOI: 10.1037//1076-898X.7.4.308

Mental Models and Logical Reasoning Problems in the GRE

Yingrui Yang
Rensselaer Polytechnic Institute

P. N. Johnson-Laird
Princeton University

The Graduate Record Examination (GRE) contains a class of complex reasoning tests known as *logical reasoning* problems. These problems are demanding for human reasoners and beyond the competence of any existing computer program. This article applies the mental model theory of reasoning to the analysis of these problems. It predicts 3 main causes of difficulty, which were corroborated by the results of 4 experiments: the nature of the logical task (Experiment 1), the set of foils (Experiment 2), and the nature of the conclusions (Experiments 3 and 4). This article shows how these factors can be applied to the design of new problems.

Most psychological studies of reasoning concern deductions that are logically straightforward, and even those problems that are difficult for human reasoners are easy for computer reasoning programs (see, e.g., Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000; Yang & Johnson-Laird, 2000). Certain inferential problems, however, are demanding for human reasoners and impossible for all current computer programs. They include a class of problems in the Graduate Record Examination (GRE), developed by the Educational Testing Service (ETS) to predict performance in graduate school. The examination has sections measuring mathematical, verbal, and analytical ability. The analytical section contains two sorts of problems, known respectively as *analytical* problems and *logical reasoning* problems. Here is an example of a logical reasoning (LR) problem:

> Children born blind or deaf and blind begin social smiling on roughly the same schedule as most children, by about three months of age.
> The information above provides evidence to support which of the following hypotheses:
> A. For babies the survival advantage of smiling consists in bonding the caregiver to the infant.
> B. Babies do not smile when no one else is present.
> C. The smiling response depends on an inborn trait determining a certain pattern of development.
> D. Smiling between persons basically signals a mutual lack of aggressive intent.
> E. When a baby begins smiling, its caregivers begin responding to it as they would to a person in conversation.

This LR problem is relatively easy, as readers should check for themselves. However, no current computer program can take such problems verbatim and reason its way through to the correct conclusion. Programs are subject to three main difficulties: the extraction of the logical structure of the problems, the variety of the inferential tasks posed by the problems, and the use of general knowledge—often in subtle ways—to solve the problems. The correct answer to the problem is Option C, the hypothesis supported by the information in the text. The information in the text implies Option C, as the following argument shows:

1. The smiling response is a behavior. (general knowledge)
2. Any behavior is an inborn trait or learned. (general knowledge)
3. Children born blind cannot see a caregiver smile. (general knowledge)
4. If children learn the smiling response and cannot see a caregiver smile, then they should take longer to learn the smiling response than most children. (general knowledge)
5. But, children born blind begin social smiling on roughly the same schedule as most children. (the premise)
6. Therefore, it is not the case both that children learn the smiling response and that children born blind cannot see a caregiver smile. (valid deduction from Arguments 4 and 5)
7. Therefore, it is not the case that children learn the smiling response. (valid deduction from Arguments 3 and 6)
8. Therefore, the smiling response depends on an inborn trait (valid deduction from Arguments 2 and 7).

LR problems are demanding for human reasoners, but, as data from ETS show, they differ widely in their difficulty. Each new batch of problems in the GRE, however, needs to be well distributed in difficulty to maintain a standardized and fair test. This requirement is even more important since the advent of the computerized GRE: it is designed to be an adaptive test in which the difficulty of the current item depends on the candidate's previous

Yingrui Yang, Department of Cognitive Sciences, Rensselaer Polytechnic Institute; P. N. Johnson-Laird, Department of Psychology, Princeton University.

Yingrui Yang was a visiting fellow at Princeton University and a postdoctoral fellow at Educational Testing Service (ETS). The research was also made possible, in part, by National Science Foundation Grant 0076287 to P. N. Johnson-Laird to study strategies in reasoning, and by a grant from the Graduate Record Examination board and ETS. We thank our colleagues in the Logical Reasoning group at ETS for their advice and help: Malcolm Bauer, Charles Davis, Larry Frase, Karen Kukich, Carol Tucker, and Ming-Mei Wang. We are also grateful to our colleagues at Princeton University and elsewhere for their useful suggestions: Victoria Bell, Zachary Estes, Yevgeniya Goldvarg, Hansjoerg Neth, Mary Newsome, Sergio Moreno Rios, Vladimir Sloutsky, and Jean-Baptiste van der Henst.

Correspondence concerning this article should be addressed to Yingrui Yang, Department of Cognitive Sciences, Carnegie Building, Rensselaer Polytechnic Institute, Troy, New York 12180-3590. Electronic mail may be sent to yangyri@rpi.edu.

performance. The original method used at ETS to determine the difficulty of a new problem is to try out the item in a so-called pretest: each real test includes a certain percentage of new items that do not count in the assessment of an individual's GRE performance. The relative difficulty of each new item is estimated by comparing its results with those from problems of a known difficulty established in previous tests. Not only is this method costly, but its validity is questionable (e.g., an examinee could spend more time on a pretest item). And once an item has been used in a computer pretest, it is, in effect, no longer secure: examinees can describe it to other potential candidates. Our research was motivated accordingly by two concerns. First, there was an urgent need to develop practical guidelines to help to predict and to manipulate the difficulty of LR problems. We therefore worked in collaboration with colleagues and item writers at ETS. Second, there was a major theoretical goal of explaining what determines the difficulty of LR problems. Our aim was to apply our results to the design of new LR problems.

We began our study with an analysis of a representative set of 120 LR problems that are in the public domain. ETS made available to us the results from tests with these problems. What all the problems have in common is a basis in real-life examples and a three-part structure: an initial text, a sentence that poses a task, and a set of five options containing one correct answer and four incorrect foils. The problems fall into various categories in terms of the following tasks they pose:

1. Problems of identifying which option can be inferred from the text, as in the preceding problem about the smiling response.
2. Problems of identifying the best hypothesis that accounts for the information in the text. The correct option is a premise from which the conclusion in the text follows, and so these problems are merely the converse of those in Category 1: they switch around the respective logical roles of text and option. We henceforth subsume these problems under Category 1. Both demand a grasp of an inferential implication.
3. Problems of identifying a missing premise in the argument of the text (e.g., "The argument above requires at least one additional premise. Which of the following could be such a required premise?")
4. Problems of identifying a weakness in an argument for a conclusion stated in the text (e.g., "Which of the following, if true, would most seriously weaken the conclusion drawn above?")
5. The sample of 120 problems included 1 problem in which the task was to determine the logical relation between two propositions in the text: "Which of the following, if true, would best resolve the apparent contradiction between statement I and statement II?"

Because such problems are rarely used in the GRE, we say no more about them.

In summary, three main categories of problems concern us: first, *inferential* problems in which individuals have to determine which option is a conclusion the text implies, or which option implies a conclusion in the text; second, *missing-premise* problems in which individuals have to determine which option states a missing premise in the text; and, third, *weakness* problems in which individuals have to identify which option states a weakness in an argument in the text.

The statement of a particular task varies from one problem to another. In some cases, a problem is stated in negative terms. For example, an inferential problem was stated in these terms: "If the position expressed is correct, then each of the following can be true EXCEPT." In other words, the task is to identify the option that is inconsistent with the text, that is, its negation follows from the text. This sort of negative task is likely to contribute to the difficulty of a problem, because negation is a well-known source of difficulty in comprehension and inference (see e.g., Wason & Johnson-Laird, 1972).

How do people solve LR problems, and what makes them difficult? No one knows the answers to these questions, and so our strategy was to start with the theory of mental models, which had been successful in predicting the difficulty of deductions and led to the discovery of some new phenomena (see e.g., Johnson-Laird & Byrne, 1991). We made the initial assumption that each of the three main components of an LR problem was likely to be a potential cause of difficulty: the text, the task, and the set of options.

In the case of the text, an index of difficulty is merely to read the text to prepare oneself to answer some as yet, unknown question. The text for the smiling problem is easy to understand: Children born blind or deaf and blind begin social smiling on roughly the same schedule as most children, by about three months of age.

Texts from other problems, however, are manifestly harder to understand:

> Two hundred corporations with net incomes of more than $122 million apiece accounted for 77 percent of total corporate gifts to the United States higher education in 1985. That year, 26 percent of total corporate gifts to United States higher education came from 14 Japanese corporations, each of which received income from 27 or more countries.

The model theory suggests a number of factors that influence the comprehension of discourse (see Johnson-Laird, 1983), but perhaps the most important is the ease of constructing appropriate models of the possibilities implied by the text. This factor, in turn, depends, in part, on the particular task posed in the question.

The three main sorts of task—inferential, missing premise, and weakness of argument—may differ in their intrinsic difficulty. It is impossible to compare all three tasks, because the nature of the task is inevitably confounded with its content. However, a comparison between inferential and missing-premise tasks is feasible. And the model theory predicts a difference in difficulty between them: It should be easier to identify a conclusion than to identify a missing premise. Consider a simple illustrative example, such as an inference of the following form:

A or B, or both.
Not A.
∴ B.

According to the model theory, the task of drawing an inference calls for constructing mental models of each of the possibilities consistent with the premises (Johnson-Laird & Byrne, 1991). In other words, each mental model represents a possibility. In fact, these premises have only a single model,

$\neg A \quad B$

where '¬' denotes negation. Reasoners can then determine which of the five options is true in the models (i.e., an option of the form, B). In contrast, a missing-premise problem has a text that is an inference and its conclusion, but without an essential premise:

> A or B, or both.
>
> Therefore, B.

Reasoners can construct the models of the disjunctive premise,

```
A
   B
A  B
```

where each row is a model of a different possibility. Reasoners can now try to determine how the models should be modified to yield the conclusion, which is *B*. One such modification is to eliminate the first model. The next step is to formulate a premise that will do the job, that is, a premise that negates the model but leaves the other models intact (e.g., If A then B). The options in the problem, however, may not contain this premise. Another modification to the models is to eliminate the first and third models. Again, this step calls for working out a premise that will do the job (i.e., not A), and checking it against the options. Hence, the task of identifying a missing premise is more complicated than the task of identifying a conclusion. The same prediction is likely to follow from theories of reasoning based on formal rules (see Braine & O'Brien, 1998; Rips, 1994). Reasoners must examine the relation between the premises and the conclusion, try to figure out what information is needed for an inference from the premises to the conclusion, and then check whether this information is among the options.

The set of options should also affect the difficulty of a problem. In the case of an inferential problem, for example, we can classify options into four sorts of categories. First, there is the valid option, which is the correct response that follows from the text (and general knowledge). Second, a foil can be consistent with the text (i.e., it may be true, but it isn't necessarily true given the text). Third, a foil can be inconsistent with the text (i.e., it is false given the text). Fourth, a foil can be irrelevant given the text (i.e., the text has no bearing on its truth or falsity, because the option introduces matters extraneous to the text). Few LR problems are likely to have options of all four sorts, and so to illustrate them we have taken an actual LR problem and introduced some new foils. The text and task are as follows:

> Computer programs are unusual in that they are virtually the only products that have been protected both by patent and by copyright. Patents protect the idea behind an innovation, whereas copyrights protect the expression of that idea. However, in order to win either protection, the idea must be clearly distinguished from its expression. Which of the following can be properly concluded from the statements given?

We state here four options to illustrate the different categories:

1. The idea behind some computer programs can be distinguished from the expression of that idea. (A valid option.)
2. Most programs are patented and copyrighted. (A consistent option.)
3. The idea behind a program can never be clearly separated from its expression. (An inconsistent option.)
4. Computer viruses cannot be patented. (An irrelevant option.)

All the foils in the problem about the smiling response are irrelevant, and the problem is easy. Indeed, if all the foils are irrelevant or inconsistent, a problem is likely to be easy. The model theory, however, makes a subtler prediction that derives from its treatment of invalid conclusions. The theory postulates that reasoners develop a variety of strategies for reasoning (Bucciarelli & Johnson-Laird, 1999; Johnson-Laird, Savary, & Bucciarelli, 1999). A conclusion can be shown to be invalid if there is no model of the premises in which it holds: it is inconsistent with the premises. However, a conclusion can also be shown to be invalid if there is a counterexample to it: it is consistent with the premises, but there is at least one model of them—the counterexample—in which the conclusion is false. In addition, the theory predicts that it should be easier to evaluate inconsistent conclusions that are invalid than consistent conclusions that are invalid. Reasoners are liable to reject a conclusion that does not match any of the models of the premises, but they may be tempted to accept a conclusion that matches one or more of the models of the premises. The only way that they can reject the latter conclusion is to check whether it holds in all the models. In this way, they may discover a counterexample to the conclusion. As an example, consider the case of a text consisting of a premise of the following form:

> A or B, or both

and two putative conclusions (or foils):

> Not-A and not-B.

and:

> Not-A and B.

The premise has the mental models:

```
A
   B
A  B
```

The first foil is not consistent with any of these models, and so it should be easy to reject. However, the second foil is consistent with one of the models, and so reasoners may be tempted to accept the foil. Reasoners often construct only one model of the premises (see e.g., Bauer & Johnson-Laird, 1993), and so they should immediately reject inconsistent conclusions, but will easily confuse a consistent conclusion, which is true in at least one model of the premises, with a necessary conclusion, which is true in all the models of the premises. A corollary, of course, is that inconsistent foils in GRE problems should be easier to assess than consistent foils.

Is there any theory that would predict otherwise? In fact, current formal rule theories do not make this prediction. These theories run parallel with the model theory in many respects, especially in their ability to yield valid conclusions (see e.g., Stenning & Yule, 1997). But, the two sorts of theory do diverge in their empirical consequences in several ways (see Johnson-Laird, 2001). One major divergence is in how the two theories establish invalidity. The late Jon Barwise first drew attention to this point (see Barwise, 1993), which we illustrate for Rips' (1994) formal rule theory. Rips' theory postulates a single deterministic strategy in evaluating the validity of a putative conclusion: it searches exhaustively through all possible formal derivations for a proof of the conclusion from the premises. The search halts either when it yields a proof or when an exhaustive search has failed to discover one. Hence, the only way to establish that a conclusion is invalid is to fail to find a formal proof for it (see Rips, 1994, p. 104). The theory therefore cannot distinguish between invalid conclusions consistent with the premises and invalid conclusions *inconsistent with the premises*. In both cases, their invalidity is established by searching for a proof and failing to find one. The model theory, as we have shown, does distinguish between the two sorts of conclusions, because possibilities play a central role in the theory. They play no role in formal rule theories. Osherson's (1976) formal rule theory contains some rules of inference for the modal operators of *necessity* and *possibility*, but, as he pointed out, his set of rules is incomplete because certain valid inferences cannot be proved with them. They do not, for instance, deal with the example above on the basis of the inclusive disjunction. This argument, however, does not imply that formal rule theories cannot, in principle, predict a difference between consistent and inconsistent invalid conclusions. It may be possible to reformulate the current theories so that they draw the distinction and make the same prediction as the model theory.

Other factors are likely to influence the difficulty of LR problems, and we return to them later, but we now describe a series of experiments that were designed to test the model theory's predictions.

Experiment 1: The Nature of the Inferential Task

In Experiment 1, we tested the prediction that inferential problems should be easier than missing-premise problems. We show that the prediction exploited the fact that an inferential problem can be converted into a missing-premise problem, and a missing-premise problem can be converted into an inferential problem. Hence, both sorts of problems can have the same content, although it is distributed differently between the text and the options.

Method

Design. The participants acted as their own controls and carried out four sorts of problems, with each having only a single option to be evaluated, because we did not want a choice of options to vitiate a direct comparison between inferential and missing-premise problems. The four sorts of problems were: (a) inferential problems presented with a single valid conclusion, (b) inferential problems presented with a single invalid conclusion, (c) missing-premise problems presented with a single correct missing premise, and (d) missing-premise problems presented with a single incorrect missing premise. Each participant carried out 12 problems—three different problems of each of the four sorts—presented in one of four random orders. Thus, the participants encountered a particular content only once, but each content occurred equally often in the four sorts of problems in the experiment as a whole.

Materials. We selected six inferential problems and six missing-premise problems from the sample of 120 problems in the public domain. Each of them was relatively difficult according to the ETS results (i.e., performance was less than 50% correct). We constructed four versions of each of these problems. Each text was followed by a single test item for the participants to evaluate. For the six inferential problems, the valid inference version used the correct option from the original problems, and the invalid inference version used the most frequently chosen invalid foil. To construct the correct missing-premise version, we swapped the valid conclusion with a premise in the text; and to construct the incorrect missing-premise version, we swapped the invalid conclusion with a premise in the text. As examples, here are the four versions of a problem constructed from what was originally a valid inferential problem:

1. Valid inference:

> Two hundred corporations with a net income of more than $122 million apiece accounted for 77 percent of total corporate gifts to United States higher education in 1985. That year, 26 percent of total corporate gifts to United States higher education came from 14 Japanese corporations, each of which received income from 27 or more countries.
>
> Assume the statements above are true. Must the following statement also be true? One or more of the 200 corporations with more than $122 million in net income received income from 27 or more countries.

2. Invalid inference pairs the same text and statement of the task with the most frequently selected erroneous foil:

> Gifts from corporations account for more than half of the total contributions to United States higher education in 1985.

This claim is consistent with the text, but it is not a valid conclusion, because the text states only the percentage of corporate contributions.

3. The correct missing-premise version:

> Two hundred corporations with a net income of more than $122 million apiece accounted for 77 percent of total corporate gifts to United States higher education in 1985. Thus, one or more of the 200 corporations with more than $122 million in net income received income from 27 or more countries.
>
> Does the above argument necessarily hold if it assumes the following? That year, 26 percent of total corporate gifts to United States higher education came from 14 Japanese corporations, each of which received income from 27 or more countries.

4. The incorrect missing-premise version:

> Two hundred corporations with a net income of more than $122 million apiece accounted for 77 percent of total corporate gifts to United States higher education in 1985. Thus, gifts from corporations account for more than half of the total contributions to United States higher education in 1985.
>
> Does the above argument necessarily hold if it assumes the following? That year, 26 percent of total corporate gifts to United States higher education came from 14 Japanese corporations, each of which received income from 27 or more countries.

We used analogous operations to create four versions of each of the original missing-premise problems.

The resulting 48 experimental problems were divided into 4 sets, each including three different problems of the four sorts, and the participants were assigned in rotation to one set of problems.

Procedure. The participants were tested individually in a quiet room. They were given two booklets. The first booklet provided the instructions and a practice problem. The experimenter read the instructions aloud while the participants followed them in the booklet. The instructions stated that they had a set of problems to solve, which called for reasoning about brief passages, and each passage was followed by a question they should answer by circling yes or no at the bottom of the page. The participants were allowed to use paper and pencil, and they were encouraged to write or draw whatever they had in mind on the problem page during the course of solving a problem. With the practice problem, the participants first had to evaluate the putative conclusion (yes or no), and then rate its relative difficulty on a 7-point scale, from 1 (*very easy*) to 7 (*very difficult*). The participants were told that there was no time limit in the experiment, but that they were expected to make progress aggressively. After participants asked any questions about the task, they were given the second booklet with each of the 12 experimental problems on a separate page. The booklets were assembled in one of four different random orders. The experimenter recorded the time a participant spent on each problem. The participants were not told that the problems were from the GRE until the final debriefing at the end of the experiment.

Participants. Twenty undergraduates from Princeton University took part in the experiment to fulfill a course requirement. None had any previous training in logic or had participated in an experiment on reasoning.

Results and Discussion

There was no reliable difference in accuracy between the problems, but the participants responded to the inferential problems significantly faster than to the missing-premise problems (Wilcoxon signed-ranks matched-pairs test, $z = 2.88, p = .002$), and they also rated them as significantly easier (Wilcoxon test, $z = 1.66$, $p = .0485$). The results are shown in Table 1. No other differences were reliable. Even without the presence of foils, the problems were difficult, and overall performance was not much above chance. This low level of performance might well have made it impossible to observe a difference in accuracy. Nevertheless, as the model theory predicted, the missing-premise problems took the participants longer and were rated as harder than the inferential problems. Given its time limit, the actual GRE test is likely to yield the same differences in difficulty.

Experiment 2: The Effect of Foils

To examine the effects of foils on performance, Experiment 2 included only inferential LR problems, that is, those for which the task was to select the option that could be inferred from the text. As we highlighted in the introduction, the model theory predicts that the presence of a foil that is consistent with the premises in the text should make a problem more difficult. The experiment was designed to test this prediction.

Method

Design. The participants acted as their own controls and carried out three inferential problems in each of four sorts: (a) original easy problems from the GRE, (b) original difficult problems from the GRE, (c) modified easy problems in which an inconsistent foil was edited to make it consistent with the text, and (d) modified difficult problems in which a consistent (and seductive) foil was edited to make it inconsistent with the text. The participants encountered a particular content only once, but each content occurred equally often in the two versions in the experiment as a whole. The problems were presented in one of four different random orders to each participant. The model theory predicts that the effect of the experimental manipulation will be to make the originally easy problems harder and to make the originally difficult problems easier.

Materials. Six difficult inferential problems (those in the previous experiment with performance less than 50% correct), and six easy inferential problems (with performance greater than 75% correct) were selected from the pool of 120 GRE problems. Each problem was used in two versions: the original and the modified version. For each of the easy problems, we changed an inconsistent foil to make it consistent with the text. Likewise, for each of the difficult problems, we changed a consistent foil to make it inconsistent with the test. For example, an original difficult problem was as follows:

> Lobsters usually develop one smaller, cutter claw and one larger crusher claw. To show that exercise determines which claw becomes the crusher, researchers placed young lobsters in tanks and repeatedly prompted them to grab a probe with one claw—in each case always the same, randomly selected claw. In most of lobsters the grabbing claw became the crusher. But in a second, similar experiment, when lobsters were prompted to use both claws equally for grabbing, most matured with two cutter claws, even though each claw was exercised as much as the grabbing claws had been in the first experiment.
>
> Which of the following is best supported by the information above?
>
> (A). Most lobsters raised in captivity will not develop a crusher claw.
>
> (B). Exercise is not a determining factor in the development of crusher claws in lobsters.
>
> (C). Cutter claws are more effective for grabbing than are crusher claws.
>
> (D). Young lobsters usually exercise one claw more than the other.
>
> (E). Young lobsters that do not exercise either claw will nevertheless usually develop one crusher and one cutter claw.

The correct answer is (D). But (B) is a seductive foil, which is consistent with the text. Hence, the experimental version of this problem kept all other options the same, but changed (B) to an inconsistent foil:

> (B) Young lobsters do not exercise their claws.

We divided the resulting 24 problems into two sets, each consisting of the originals of three easy and three difficult problems, and modified versions of three easy and three difficult problems. We assigned a set to each participant at random.

Procedure. The participants were tested individually in a quiet room. They were given two booklets. The first booklet provided the instructions and a practice problem. The experimenter read the instructions aloud while the participants followed along in the booklet. The instructions stated that the participants had to solve a set of problems based on reasoning about brief passages, and each passage was followed by a set of five options. As

Table 1
Percentages of Correct Responses, Mean Latencies (in Minutes), and the Means of the Rated Difficulties in Experiment 1

	Inferential problems			Missing-premise problems		
Measure	% correct	Latency	Rating	% correct	Latency	Rating
Valid conclusions or correct missing premises	51	2.75 (1.30)	3.20 (0.96)	61	3.21 (1.04)	3.82 (1.18)
Invalid conclusions or incorrect missing premises	63	2.84 (1.14)	3.35 (1.22)	65	3.31 (1.33)	3.49 (0.89)

Note. The means of difficulties were rated on a 7-point scale, ranging from 1 (*very easy*) to 7 (*very difficult*). Standard deviations are given in parentheses for each mean of latency and each mean of rated difficulty.

in the actual GRE, the instructions also stated that for some problems, more than one of the options could conceivably be the answer to the question. However, the participants should choose the best answer, that is, the response that most accurately and completely answered the question. They should respond by circling the relevant option. Participants were allowed to use paper and pencil, and they were encouraged to write or draw whatever they had in mind on the problem page during the course of solving a problem. The participants were also told that there was no time limit but that they were expected to make progress aggressively. After they asked any questions about the task, they were given the second booklet with each of the 12 experimental problems on a separate page. The participants were not told that the problems were from the GRE until the final debriefing at the end of the experiment.

Participants. Thirty-two Princeton University undergraduates took part in the experiment to fulfill a course requirement. None had received any previous training in formal logic or had participated in an experiment on reasoning.

Results and Discussion

Table 2 presents the overall percentages of correct responses to the four sorts of problems. As predicted, the modified easy problems were made harder, whereas the modified difficult problems were made easier, and this interaction was highly reliable (Wilcoxon test, $z = 4.25, p < .0001$). Hence, it is simple to increase the difficulty of an easy problem by introducing a consistent foil and to decrease the difficulty of a hard problem by replacing a consistent foil with an inconsistent one. As it happens, the modified problems no longer differed reliably in their difficulty (Wilcoxon test, $z = .21$, *ns.*). However, the manipulation of the foils did not make the difficult problems as easy as the original versions of the easy problems, or make the easy problems as hard as the original versions of the difficult problems. Hence, other factors must influence difficulty. Indeed, Experiment 1 showed that the difficult problems were just as difficult when there was only a single conclusion to be evaluated. We examined the effect of conclusions in the next experiment.

Table 2
Percentages of Correct Responses in Experiment 2

Problem type	Easy problems	Difficult problems
Original version	99	53
Modified version	78	70

Experiments 3 and 4: The Nature of Conclusions

When reasoners have to evaluate only a single conclusion (as in Experiment 1), the model theory predicts that the difficulty of identifying a conclusion as invalid should depend on whether it is consistent with the premises or inconsistent with them. A consistent conclusion matches at least one model of the premises and so, as we mentioned earlier, reasoners may be tempted to assume that it follows from the premises. In contrast, an inconsistent conclusion conflicts with all the models of the premises, and so its status is unequivocal. Consider, for example, the following simplified example, where the text consists of premises of the form:

A or B, or both.

If B then C

B

Each model represents a possibility, and so the models of all three premises are as follows:

```
A  B  C
   B  C
```

The putative conclusion, not-A, is invalid, but it should be relatively difficult to reject because it is consistent with the second model of the premises. In contrast, the putative conclusion, not-C, is invalid, but it should be relatively easy to reject because it is inconsistent with both of the models of the premises.

In general, valid conclusions should be easier to evaluate than invalid conclusions, and, in general, easy problems (as defined by results from the GRE) should be easier to evaluate than hard problems (as defined by results from the GRE). But according to the model theory, these main effects should be modulated by the nature of the invalid conclusions (i.e., by whether they are consistent or inconsistent with the premises). In particular, an easy invalid conclusion should be even easier to evaluate when it is inconsistent with the premises, and a hard invalid conclusion should be even harder to evaluate when it is consistent with the premises. The theory therefore predicts an interaction in this case: The difference between valid and invalid conclusions should be greater for hard problems than for easy problems. In contrast, an easy invalid conclusion should be harder to evaluate when it is consistent with the premises, and a hard invalid conclusion should be easier to evaluate when it is inconsistent with the premises. The theory therefore predicts a second interaction in this case: The difference between valid and invalid conclusions should be greater for easy problems than for hard problems. Experiment 3 tested the first of these interactions, and Experiment 4 tested the second of these interactions. The two experiments used similar designs and procedures, and so we report them together.

Method

Design. In both experiments, the participants evaluated single conclusions to easy and hard inferential problems. Half the conclusions were valid and half the conclusions were invalid. The participants acted as their own controls and carried out three problems of each of four sorts: valid easy problems, invalid easy problems, valid hard problems, and invalid hard problems. Their task was to determine whether each conclusion followed from the text. The participants encountered a particular content only once, but each content occurred equally often in valid and invalid problems in the experiment as a whole. The problems were presented in one of four different random orders to each participant.

Materials. The materials in both experiments were based on the same set of 12 problems as those in Experiment 2. We constructed two versions of each problem. In Experiment 3, the easy problems had as their putative conclusion either the original valid conclusion or an original inconsistent foil, and the difficult problems had as their putative conclusion either their original valid conclusion or an original, invalid consistent foil. Here is an example of the two versions (valid and invalid) of an easy problem:

> Most television viewers estimate how frequently a particular type of accident or crime occurs by how extensively it is discussed on television news shows.

> Television news shows report more on stories that include dramatic picture such as fires and motor vehicle accidents than they do on more common stories that have little visual drama such as bookkeeping fraud.
>
> Assume the statements above are true. Can it be properly concluded that the following is also true?
>
> (Valid conclusion) Viewers of television news shows tend to overestimate the number of fires and motor vehicle accidents that occur relative to the number of crimes of bookkeeping fraud.
>
> (Invalid foil) The usual selection of news stories for television news shows is determined by the number of news reporters available for assignment.

In Experiment 4, the materials were also based on those in Experiment 2. The easy problems had either the original conclusion (valid) or the modified foil that was consistent with the text (invalid), and the difficult problems had either the original conclusion (valid) or the modified foil that was inconsistent with the text (invalid). Here is an example of the two versions (valid and invalid) of an easy problem:

> The greater the division of labor in an economy, the greater the need for coordination. This is because increased division of labor entails a larger number of specialized products, which results in a greater burden on managers and, potentially, in a greater number of disruptions of supply and production. There is always more division of labor in market economies than in planned economies.
>
> Assume all of the statements given are true. Must the following also be true?
>
> (Valid conclusion) The need for coordination in market economies is greater than in planned economies.
>
> (Invalid foil) Disruptions of supply and production are a result of a larger number of specialized products.

In both experiments, we divided the resulting 24 problems into 2 sets, each consisting of three easy and three difficult problems with valid conclusions, and three easy and three difficult problems with invalid conclusions. We assigned a set to each participant at random.

Participants and procedure. We tested 2 separate groups of 20 Princeton University undergraduates in the two experiments. They had not studied logic or participated in an experiment on reasoning. The procedure was the same as in Experiment 1.

Results and Discussion

Table 3 presents the results of Experiment 3. Overall, as predicted, the easy problems were reliably easier than the difficult problems on all three measures (Wilcoxon test, $z \geq 3.0, p \leq .0013$, in all three cases). Likewise, the valid problems yielded a greater percentage of correct responses than the invalid ones (Wilcoxon test, $n = 15, T = 116, p = .0002$) and were rated as more difficult (Wilcoxon test, $z = 2.6, p = .0047$). However, the increase in difficulty from valid to invalid conclusions was not greater for the hard problems than for the easy problems. What happened was the invalid difficult problems were, as the model theory predicts, very hard: performance sank to a level of accuracy no better than chance. This "floor" effect made it impossible to observe the predicted interaction.

Table 4 presents the results of Experiment 4. The effect of the experimental manipulation was striking. Overall, the difference between the easy and difficult problems almost disappeared. Indeed, the percentages of correct responses suggest—especially in the case of the invalid conclusions, as predicted—that the easy problems have now become more difficult than the hard problems, where easy and hard are defined in terms of GRE results. In fact, the only significant main effect is that the latencies are shorter for the easy problems than for the hard problems (Wilcoxon test, $z = 3.85, p < .0001$). The predicted interaction was strongly corroborated: The increase in difficulty from valid to invalid conclusions was reliably larger for the easy problems than for the difficult problems (Wilcoxon test, $z = 2.03, p = .0212$, on all three measures). One puzzle, however, is why the difference between easy and hard valid conclusions is apparent in Experiment 3, but seems to have disappeared in Experiment 4. The answer, as we discuss below, is likely to be in the different strategies that the participants developed for the two experiments.

General Discussion

The model theory predicts that three main factors should affect the difficulty of GRE, LR problems, and our experiments have corroborated their effects. The first factor is the nature of the task. Experiment 1 confirmed that *inferential problems are easier than missing-premise problems.* Hence, the nature of the logical task can affect performance. However, we have not investigated the relative difficulty of LR problems that require individuals to identify the weakness in an argument. It may not be possible to do so in an unconfounded way, because such a problem is bound to differ in content from those for the inferential and missing-premise problems.

The second factor is the nature of the foils. The model theory predicts that it should be easier to reject a foil that is inconsistent with a text than a foil that is consistent with the text. The reason is that *a consistent foil, by definition, corresponds to one of the possibilities admitted by the premises.* If this possibility is represented in a mental model of the premises, then reasoners may well infer that it is valid. Experiment 2 corroborated this prediction,

Table 3
Percentages of Correct Responses, Mean Latencies (in Minutes), and Means of Rated Difficulties for the Four Sorts of Problems in Experiment 3

	Easy problems			Difficult problems		
Measure	% correct	Latency	Rating	% correct	Latency	Rating
Valid conclusion	100	1.78 (0.65)	2.08 (0.90)	75	3.22 (1.42)	3.93 (1.37)
Invalid conclusion	83	1.82 (0.81)	2.82 (1.03)	58	3.20 (1.55)	4.28 (1.84)

Note. Standard deviations are given in parentheses for each mean of latency and each mean of rated difficulty.

Table 4
Percentages of Correct Responses, Mean Latencies (in Minutes), and Means of Rated Difficulties for the Four Sorts of Problems in Experiment 4

	Easy problems			Difficult problems		
Measure	% correct	Latency	Rating	% correct	Latency	Rating
Valid conclusion	83	2.04 (0.92)	2.93 (1.35)	92	2.94 (1.66)	3.67 (1.83)
Invalid conclusion	53	2.27 (1.03)	3.47 (1.52)	90	2.77 (1.60)	3.50 (1.49)

Note. Standard deviations are given in parentheses for each mean of latency and each mean of rated difficulty.

showing that an easy inferential problem can be made harder by introducing a foil that is consistent with the text, and a difficult problem can be made easier by introducing a foil that is inconsistent with the text. The nature of the options changes in the case of missing-premise and weakness problems. Thus, in the case of a missing-premise problem, the correct option is indeed a missing premise, and an option that follows from the admittedly incomplete text is a foil. And, in the case of a weakness problem, the correct option may be inconsistent with the text. In future studies, we plan to examine the influence of different sorts of foils on missing-premise and weakness problems.

The third factor is the relation between the text and the conclusion. Experiments 3 and 4 showed that easy problems are easier than difficult problems when there are no foils, and the task is merely to evaluate a valid conclusion. The difficulty of rejecting an invalid conclusion depends on whether it is consistent or inconsistent with the text. In Experiment 3, consistent conclusions reduced performance with difficult problems to the level of chance. In Experiment 4, they reduced performance with easy problems to the level of chance. Hence, in this experiment, the difference between valid and invalid problems was large for the easy problems, but disappeared for the difficult problems.

When individuals tackle a series of deductions, they tend to develop a strategy for coping with them (see e.g., Bucciarelli & Johnson-Laird, 1999; van der Henst, Yang, & Johnson-Laird, 2001). Different individuals develop different strategies, for example, some reasoners tend to follow-up the consequences of suppositions, whereas others draw diagrams of the possibilities consistent with the premises. We surmise that individuals are also likely to develop strategies for coping with LR problems if they have not already been taught a strategy. The results of Experiments 3 and 4 suggest that participants may have developed a strategy based on conclusions that were inconsistent with the texts. The strategy was to respond *no* to any such problem, which is correct, and to respond *yes* to any other problem. The consequence in Experiment 3 was that performance with the invalid difficult problems, which had conclusions consistent with the texts, fell to a chance level. However, the consequence was more striking in Experiment 4. The difficult problems became easy because it was simple to determine that a conclusion was invalid—it was inconsistent with the premises. But performance with the invalid conclusions to the easy problems, which are consistent with the text, dropped to chance levels. If reasoners did develop such a strategy during the course of the experiments, then these strategies were exquisitely tuned to the exigencies of the problems.

Our results have immediate practical applications for the design of new LR problems in the GRE, and perhaps for the design of analogous problems in the Scholastic Assessment Test (SAT). Our results enable the devisers of tests to carry out simple manipulations to alter the level of difficulty of a problem. Consider, for instance, an LR inferential problem. The test developer can increase its difficulty either by changing the problem to that of a missing premise or by introducing a foil that is consistent with the text. The developer can reduce its difficulty by eliminating any foil that is consistent with the premises or by introducing a foil that is inconsistent with them. Now consider an LR missing-premise problem. The test developer can reduce its difficulty by changing the task to an inferential one or by changing any consistent foils so that they become inconsistent with the text. For an inferential problem of identifying what does not follow from the text, the task can be made easier by ensuring that the correct option is inconsistent with the text, and it can be made more difficult by ensuring that the correct option is consistent with the text (though it does not follow from it).

Finally, our results have theoretical implications. They show that the theory of mental models can account for at least three major factors that influence the difficulty of LR problems in the GRE: the nature of the task, the nature of the foils, and the nature of the conclusions. Our experiments have corroborated their predicted effects. Theories based on formal rules can probably account for the difference in difficulty between inferential and missing-premise problems. But current formal rule theories make no use of possibilities (see e.g., Braine & O'Brien, 1998; Rips, 1994), and so they make no predictions about the effects on inferential problems of foils that are consistent with the premises. Likewise, these theories assume that invalidity is established only by a failure to find a proof of a conclusion, and so they make no predictions about the differences in difficulty of evaluating invalid conclusions that are consistent, or inconsistent, with the premises. The text itself is also likely to contribute to the difficulty of a problem, although its role is bound to depend on the nature of the task. We conjecture that the ease of constructing the set of possibilities—the set of models—corresponding to the text lies at the heart of this factor, as does the complexity of the relational structure of these models, but the elucidation of these factors remains to be investigated in future work.[1]

[1] We thank Ray Nickerson and three anonymous reviewers for their helpful critiques of earlier versions of this article.

References

Barwise, J. (1993). Everyday reasoning and logical inference. *Behavioral and Brain Sciences, 16,* 337–338.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science, 4,* 372–378.

Braine, M. D. S., & O'Brien, D. P. (Eds). (1998). *Mental logic.* Mahwah, NJ: Erlbaum.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science, 23,* 247–303.

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. (2001). Mental models and deduction. *Trends in Cognitive Science, 5,* 434–442.

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

Johnson-Laird, P. N., Legrenzi, P., Girotto, P., & Legrenzi, M. S. (2000). Illusions in reasoning about consistency. *Science, 288,* 531–532.

Johnson-Laird, P. N., Savary, F., & Bucciarelli, M. (1999). Strategies and tactics in reasoning. In W. S. Schaeken, G. De Vooght, A. Vandierendonck, & G. d'Ydewalle (Eds.), *Deductive reasoning and strategies* (pp. 209–240). Mahwah, NJ: Erlbaum.

Osherson, D. N. (1976). *Logical abilities in children, Vol. 4: Reasoning and concepts.* Hillsdale, NJ: Erlbaum.

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

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

van der Henst, J. B., Yang, Y., & Johnson-Laird, P. N. (2001). *Strategies in sentential reasoning.* Manuscript submitted for publication.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content.* Cambridge, MA: Harvard University Press.

Yang, Y., & Johnson-Laird, P. N. (2000). Illusions in quantified reasoning: How to make the impossible seem possible, and vice versa. *Memory & Cognition, 28,* 452–465.

Received November 17, 2000
Revision received September 25, 2001
Accepted September 25, 2001 ■

New Editors Appointed, 2003–2008

The Publications and Communications Board of the American Psychological Association announces the appointment of five new editors for 6-year terms beginning in 2003.

As of January 1, 2002, manuscripts should be directed as follows:

- For the *Journal of Applied Psychology,* submit manuscripts to **Sheldon Zedeck, PhD,** Department of Psychology, University of California, Berkeley, CA 94720-1650.

- For the *Journal of Educational Psychology,* submit manuscripts to **Karen R. Harris, EdD,** Department of Special Education, Benjamin Building, University of Maryland, College Park, MD 20742.

- For the *Journal of Consulting and Clinical Psychology,* submit manuscripts to **Lizette Peterson, PhD,** Department of Psychological Sciences, 210 McAlester Hall, University of Missouri—Columbia, Columbia, MO 65211.

- For the *Journal of Personality and Social Psychology: Interpersonal Relations and Group Processes,* submit manuscripts to **John F. Dovidio, PhD,** Department of Psychology, Colgate University, Hamilton, NY 13346.

- For *Psychological Bulletin,* submit manuscripts to **Harris M. Cooper, PhD,** Department of Psychological Sciences, 210 McAlester Hall, University of Missouri—Columbia, Columbia, MO 65211.

Manuscript submission patterns make the precise date of completion of the 2002 volumes uncertain. Current editors, Kevin R. Murphy, PhD, Michael Pressley, PhD, Philip C. Kendall, PhD, Chester A. Insko, PhD, and Nancy Eisenberg, PhD, respectively, will receive and consider manuscripts through December 31, 2001. Should 2002 volumes be completed before that date, manuscripts will be redirected to the new editors for consideration in 2003 volumes.