# A study of complex reasoning: The case of GRE 'logical' problems

**Yingrui Yang** (yingruiy@phoenix.princeton.edu)
Department of Psychology, Princeton University,
Green Hall, Princeton, NJ 08544, USA

**P.N. Johnson-Laird** (phil@clarity.princeton.edu)
Department of Psychology, Princeton University,
Green Hall, Princeton, NJ 08544, USA

**Abstract**

Complex reasoning, such as that elicited by GRE 'logical' reasoning problems, is demanding for human reasoners and beyond the competence of any existing computer program. We report four experiments carried out to investigate the question of what makes these problems difficult. The experiments established three causes of difficulty: the nature of the logical task (Experiment 1), the nature of the incorrect foils (Experiment 2), and the nature of the correct conclusions (Experiments 3 and 4).

## Introduction

Most psychological studies of reasoning concern deductions that are logically straightforward. Even those deductions that are difficult for human reasoners are easy for computer reasoning programs. 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 Educational Testing Service to predict performance in graduate school. The examination has sections measuring mathematical, verbal, and analytical ability. The analytical section contains two sorts of problem, which are known respectively as analytical problems and logical problems. Here is an example of a GRE logical reasoning 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 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 problem is easy, as readers may check for themselves. (The correct answer is below). Yet, no current computer program can take such problems verbatim and reason its way through to the correct conclusion. There are at least three difficulties: the extraction of the logical structure of the problems, the variety of the inferential tasks posed by the problems, and their reliance often in subtle ways on general knowledge. Some GRE logical reasoning problems are much harder than this example, but 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 the task, and a set of five options containing one correct answer and four incorrect foils.

No-one knows what makes an GRE logical reasoning problem difficult for human reasoners. Our aim in the present paper is to begin to answer this question. Our studies relied on the 120 GRE logical problems that are in the public domain. The correct answer to the problem above is option C. It depends on the following argument: If blind children start to smile at the same point in their development as

sighted children, then smiling is not learned, because it could be learned only from seeing a caregiver smile. If smiling is not learned, it must depend on an inborn disposition. This problem calls for an inference, but other problems pose other tasks. We begin our investigation with a study of these tasks.

### The nature of the inferential task: Experiment 1

GRE logical problems fall into four principal categories in terms of the tasks they pose: 1. Problems of identifying which option can be inferred from the text, 2. Problems of identifying a missing premise in the argument of the text, 3. Problems of identifying a weakness in the argument in the text, and 4. Problems of identifying the logical relation between two propositions in the text. These tasks probably differ in their intrinsic difficulty, but it is impossible to make a general comparison, because the nature of the task is inevitably confounded with its content. One comparison, however, is feasible, and it concerns a difference that the mental model theory predicts. 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 theory of mental models (see e.g. Johnson-Laird and Byrne, 1991), the task of drawing an inference calls for constructing models of each of the possibilities consistent with the premises. In fact, these premises have only a single model:

¬ A B

where "¬" denotes negation. Reasoners can then determine which option is true in the models, i.e. an option of the form, B. In contrast, a missing-premise problem has a text which is an inference with a missing premise:

A or B, or both.
∴ B.

Reasoners can construct the models of the disjunctive premise:

A ¬ B
¬ A B
A B

where each row is a model of a different possibility. Reasoners can now try to work out how the models should be modified in order to yield the conclusion: 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. 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.

Experiment 1 tested the prediction that inferential problems should be easier than missing-premise problems. It used problems with an identical content, consisting of a text followed by a single test item for the participants to evaluate. The problems were based on six inferential problems and six missing-premise problems from the sample of 120 GRE problems. All 12 problems were difficult, as performance with them in the GRE showed. We constructed four versions of each of the problems: a valid inference, an invalid inference, a correct missing premise, and an incorrect missing premise. The valid conclusion was the original correct option, the invalid conclusion was the most frequently chosen foil in the original item, and so on. The resulting 48 experimental problems were divided into four sets, each including three different problems the four sorts, and the participants carried out all the problems in a set. Thus, the participants encountered a particular content only once, and carried out three problems of the four different sorts, but each content occurred equally often in the three sorts of problem in the experiment as a whole. The participants were allowed to use paper and pencil, and were encouraged to write or draw whatever they had in mind on the problem page during the course of solving a problem. After they had evaluated the putative conclusion, they rated its difficulty on a 7-point scale. We tested 20 Princeton undergraduates individually.

Table 1: The percentages of correct responses, the mean latencies (in minutes), and the means of the rated difficulties to the four sorts of problems in Experiment 1.

| | Inferential problems | | |
|---|---|---|---|
| | % Correct | Latency | Rating |
| Valid conclusion | 51 | 2.75 | 3.20 |
| Invalid conclusion | 63 | 2.84 | 3.35 |
| | Missing-premise problems | | |
| | % Correct | Latency | Rating |
| Valid conclusion | 61 | 3.21 | 3.82 |
| Invalid conclusion | 65 | 3.31 | 3.49 |

The results are shown in Table 1. 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 ($z=2.88$, $p<.002$), and they also rated them as significantly easier ($z=1.66$, $p<.05$). (Here and throughout the paper, the tests of significance were Wilcoxon signed-ranks matched-pairs tests.) No other differences were reliable. The results accordingly corroborated our prediction that it should be easier to evaluate a putative conclusion than a putative missing premise.

## The effect of foils: Experiment 2

Experiment 2 concerned only those GRE logical reasoning problems in which the task was to select the option that could be inferred from the text. A salient source of difficulty should be the set of five options from which the testees select their choice. We can classify options into those that are: 1. valid, i.e. they follow from the text (and general knowledge); 2. consistent with the text, i.e. they may be true given the text, but they do not follow from it; 3. inconsistent, i.e. they are false given the text; and 4. irrelevant, i.e. whether true or false, they are consistent with the text. Readers should note that all the foils in the sample problem in the Introduction are irrelevant. Indeed, if all the foils are irrelevant or inconsistent, a problem is likely to be easy. But, the presence of a consistent foil should render a problem more difficult. Reasoners often fail to construct all the models of the premises, and in this case they may easily confuse a consistent conclusion, which is true in some of the models of the premises, with a necessary conclusion, which must be true in all the models of the premises (Johnson-Laird and Byrne, 1991). We therefore used six difficult inferential problems (those in the previous experiment) and six easy inferential problems from the pool of 120 GRE problems. For each of the difficult problems, we changed the most seductive foil (as shown by the results from the GRE test) to make it inconsistent with the text. Likewise, for each of the easy problems, we changed an inconsistent foil to make it consistent with the text. We divided the resulting 24 problems into two sets, each consisting of the original versions of three difficult and three easy problems, and modified versions of the other six problems. We assigned a set to each participant at random, so that he or she saw a text only once. We tested 32 Princeton undergraduates individually. They were allowed to take as much time as they needed for each problem.

Table 2: The percentages of correct responses in Experiment 2

| | Easy problems / Difficult problems | |
|---|---|---|
| Original version | 99 | 53 |
| Modified version | 78 | 70 |

Table 2 presents the percentages of correct responses to the four sorts of problem. As predicted, the modified easy problems were made harder whereas the modified difficult problems were made easier, and the interaction was highly reliable ($z = 4.25$, $p < .0001$). 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 introducing an inconsistent foil. As it happens, the modified problems no longer differed reliably in difficulty ($z = .21$, $p > .4$). But, the manipulation of the foils was unable to make the difficult problems as easy as the original versions of the easy problems, or to make the easy problems as hard as the original versions of the difficulty problems. Hence, other factors must influence difficulty. Indeed, Experiment 1 showed that the difficult problems were just as difficult when there were no foils. We examined the effect of conclusions in the next experiment.

## The nature of conclusions: Experiment 3

In this experiment, each text was presented with only a single putative conclusion,

and the task was to determine whether or not it followed from the text. The materials were based on the same set of problems as those in the previous experiment. The easy problems had as their putative conclusion either the original conclusion (valid) or the original inconsistent foil (invalid), and the difficult problems had as their putative conclusion either their original conclusion (valid) or the original consistent foil (invalid). The participants were allowed to use paper and pencil, and the procedure was the same as Experiment 1. We tested 20 Princeton undergraduates individually.

Table 3: The percentages of correct responses, the mean latencies (in minutes), and the means of the rated difficulties to the four sorts of problems in Experiment 3.

| | Easy problems | | |
|---|---|---|---|
| | % Correct | Latency | Rating |
| Valid conclusions | 100 | 1.78 | 2.08 |
| Invalid conclusions | 83 | 1.82 | 2.82 |
| | Difficult problems | | |
| | % Correct | Latency | Rating |
| Valid conclusions | 75 | 3.22 | 3.93 |
| Invalid conclusions | 58 | 3.20 | 4.28 |

Table 3 presents the results of the experiment. Overall, the easy problems were reliably easier than the difficult problems on all three measures ($z \geq 3.0$, $p \leq .002$, in all three cases). The valid problems yielded a greater percentage of correct responses than the invalid ones ($n = 15$, $c > 116$, $p < .0002$) and were rated as more difficult ($z = 2.6$, $p < .005$). The invalid easy problems depended on a conclusion that was inconsistent with the text, and the invalid difficult problems depended on a conclusion that was consistent with the text. Hence, the invalid difficult problems should show a greater increase in difficulty than the invalid easy problems. This prediction was not reliable, perhaps because the invalid difficult problems sank to a level of accuracy that was no better than chance. This 'floor' effect may have vitiated the latency and rating measures.

### Experiment 4

In this experiment, each text was again presented with only a single putative conclusion, and the task was to determine whether or not it followed from the text. The materials were based on the same set of 24 problems as 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. The procedure was the same as the previous experiment. We tested 20 Princeton undergraduates individually.

Table 4: The percentages of correct responses, the mean latencies (in minutes), and the means of the rated difficulties to the four sorts of problems in Experiment 4.

| | Easy problems | | |
|---|---|---|---|
| | % Correct | Latency | Rating |
| Valid conclusion | 83 | 2.04 | 2.93 |
| Invalid conclusion | 53 | 2.27 | 3.47 |
| | Difficult Problems | | |
| | % Correct | Latency | Rating |
| Valid conclusion | 92 | 2.94 | 3.67 |
| Invalid conclusion | 90 | 2.77 | 3.50 |

Table 4 presents the results of the experiment. At first sight, the effect of the experimental manipulation was extraordinary. Overall, the difference between the easy and difficult problems almost disappeared: if anything, the accuracies switched around, though not reliably, perhaps because performance with the invalid easy problems was at chance. Only the latencies show a significant advantage for the easy problems ($z = 3.85$, $p < .0001$). What is striking is the interaction: on every measure, the easy problems show an advantage for the valid conclusion, whereas the difference disappears for the difficult problems (the interaction was reliable, $z \geq 2.03$, $p \leq .05$ on all three measures).

### General Discussion

Our experiments have shown the existence of three factors that affect the difficulty of GRE logical problems. The first factor is the nature of the task. Experiment 1 corroborated the model theory's prediction that inferential problems are easier than missing-premise problems. Hence, the nature of the logical task affects performance. A second factor is the nature of the foils. Experiment 2 showed that an easy 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. A third factor is the relation between the text and the correct conclusion. Experiment 3 showed that easy problems are easier than difficult problems even when there are no foils and the task is merely to evaluate a single putative conclusion. When we consider the results of Experiment 3 and 4 together, a striking and unexpected phenomenon emerges. The participants developed a strategy based on problems with invalid conclusions that were inconsistent with the texts. They showed a bias towards responding "no" to any such problem, which is the correct response. This bias, however, often led them to often respond "yes" to any other problem. The result in Experiment 3 was that performance with the invalid difficult problems, which have conclusions that are consistent with the texts, fell to a chance level. The effect was more dramatic 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 – the participants responded 'yes', because the conclusion was not inconsistent with the text. The moral of these studies is that reasoners develop particular strategies during the course of experiments (see also Johnson-Laird, Savary, and Bucciarelli, 1999), and these strategies can be exquisitely tuned to the exigencies of the problems.

Complex reasoning problems, such as the logical problems in the GRE examination, vary in their difficulty. We have succeeded in isolating three causes of difficulty – the nature of the task, the nature of the foils, and the nature of the conclusions. A task for the future is to determine how the logical structure of the text, taken in conjunction with the correct response, influences difficulty.

**Acknowledgements**

Yingrui Yang is a visiting fellow at Princeton University and an E.T.S. post-doctoral fellow. We are grateful to colleagues in the Logical Reasoning group at E.T.S. for their advice and help: Malcolm Bauer, Charles Davis, Larry Frase, Karen Kukich, Carol Tucker, and Ming-Mei Wang. We also thank our colleagues at Princeton 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.

**References**

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

Johnson-Laird, P.N., Savary, F., and Bucciarelli, M. (1999) Strategies and tactics in reasoning. In W.S. Schaeken, A. Vandierendonck, G. De Vooght, & G. d'Ydewalle (Eds.) Deductive Reasoning and Strategies. Mahwah, NJ: Lawrence Erlbaum Associates. In press.