EUROPEAN JOURNAL OF COGNITIVE PSYCHOLOGY, 1994, 6 (3) 263–285

Debiasing by Instruction: The Case of Belief Bias

J.St.B.T. Evans, S.E. Newstead
Department of Psychology, University of Plymouth, Plymouth, UK

J.L. Allen
School of Social Sciences, Trinity and All Saints College, Leeds, UK

P. Pollard
Department of Psychology, University of Central Lancashire, Preston, UK

The study is concerned with the question of whether robust biases in reasoning can be reduced or eliminated by verbal instruction in principles of reasoning. Three experiments are reported in which the effect of instructions upon the belief bias effect in syllogistic reasoning is investigated. Belief bias is most clearly marked by a tendency for subjects to accept invalid conclusions which are *a priori* believable. Experiment 1 attempted to replicate and extend an experiment reported by Newstead, Pollard, Evans and Allen (1992). In contrast with their experiment, it was found that belief bias was maintained despite the use of augmented instructions which emphasised the principle of logical necessity. Experiment 2 provided an exact replication of the augmented instructions condition of Newstead et al., including the presence of problems with belief-neutral conclusions. Once again, significant effects of conclusion believability were found. A third experiment examined the use of elaborated instructions which lacked specific reference to the notion of logical necessity. The use of these instructions significantly reduced the effects of belief on the reasoning observed.

Taking the current findings together with the experiment of Newstead et al., the overall conclusion is that elaborated instructions can reduce the belief bias effect in syllogistic reasoning, but not eliminate it. This conclusion is discussed with reference to (1) the practical implications for improving thinking and reasoning via verbal instruction and (2) the nature of the belief bias phenomenon.

Requests for reprints should be addressed to J.St.B.T. Evans, Department of Psychology, University of Plymouth, Plymouth, Devon PL1 5RR, UK.

INTRODUCTION

Many hundreds of experiments have been reported in the psychological literature which provide evidence of all manner of "biases" in human reasoning and judgement (see Baron, 1988; Evans, 1989; Evans, Newstead, & Byrne, 1993a). Psychologists and philosophers have argued at length over the implications that this work has for human rationality (see Manktelow & Over, 1993, for a recent collection of papers on this issue). However, it is widely believed that biases demonstrated in the laboratory are indicative of the fallibility of real-world decision making, and for this reason there is concern with the question of whether biases can be corrected and if so by what means.

This problem has been given the rather ugly name "debiasing" (Fischhoff, 1982; see also Evans, 1989, ch. 6). There are several possible approaches. One is recalibration. For example, if people are consistently overconfident, then we might get more accurate results by, say, rescaling a subjective confidence of 85% to a 70% estimate that the individual is correct in their judgement. A second approach is to replace the human judge, for example with an expert system computer program. The method of most interest to psychologists, however, is that of education and training. This divides in two: one approach is to provide feedback training in the hope of improving judgements; the other is to provide verbal instruction in relevant principles (e.g. statistical or logical) in the hope of improving performance.

In this paper, we look at the issue of whether biases can be reduced or eliminated by verbal instruction in relevant principles. Certainly, psychologists interested in the teaching of thinking are wont to regard instruction in rules and principles as a useful approach (see Baron, 1985, ch. 5; Nickerson, Perkins, & Smith, 1985). The issue is of theoretical interest as well as practical importance. Evans (1989) has argued that most reasoning and judgemental biases are caused by pre-conscious heuristics which cause attention to be focused on selected aspects of the problem information. Although subjects are capable of analytic reasoning in which they demonstrate an understanding of logical principles, this will not help them if they are attending to the wrong information. According to this view, verbal instruction will be of limited help in debiasing because instructions operate at the explicit, analytic level and not at the implicit or heuristic level.

A rather limited number of studies in the literature have attempted to debias reasoning by verbal instruction. In studies of "confirmation bias" —an alleged tendency to seek confirming evidence when testing hypotheses—studies which instruct subjects to falsify hypotheses or to test multiple hypotheses find little facilitation of performance relative to controls (see Evans, 1989, ch. 3). Apparent exceptions—such as the study of

Gorman and Gorman (1984)—occur only when subjects are instructed in the precise strategy (negative testing) which will improve performance.

There is some evidence that instructional training may be more beneficial to statistical inference than to logical reasoning. Fong, Krantz and Nisbett (1986) found that statistical reasoning about everyday problems was improved by attendance of statistics courses. The same authors found facilitation by laboratory-based training in statistical principles demonstrated either by examples or explicit statement of rules. The benefits transferred to a problem domain different from that used for training, although a subsequent study by Fong and Nisbett (reported by Holland, Holyoak, Nisbett, & Thagard, 1986) found that this transfer disappeared when testing was delayed. A parallel study of deductive reasoning by Cheng, Holyoak, Nisbett and Oliver (1986), however, found very different results. Neither attendance at logic classes nor laboratory training in abstract logical principles improved performance on a reasoning task.

It may be argued that the reasoning task used by Cheng et al.—the Wason selection task (Wason, 1966)—was an unfortunate choice. Evans (1989) and others (see Evans et al., 1993a, ch. 4) have claimed that although this task requires an understanding of conditional logic for its solution, in practice it elicits little or no reasoning from the subjects to whom it is given. Subjects may just select the cards which appear *relevant* based on pre-conscious heuristics and not engage in any kind of analytic reasoning process. The task studied in this paper—syllogistic reasoning—is a much more interesting case from this point of view, since analytic reasoning clearly also occurs on the tasks in which the bias is demonstrated. Although subjects make many errors, they do succeed in solving syllogism at above chance rates as we shall see.

Reasoning with categorical syllogisms is one of the most common paradigms used to study the psychology of deductive inference. The basic phenomena—and the many different theories of syllogistic reasoning that have been proposed—are discussed in detail by Evans et al. (1993a, chs 7 and 8). Syllogisms consist of two premises and a conclusion which link three classes. For example:

No A are B
Some C are B
Therefore, some A are not C

The two classes linked in the conclusion, A and C, are separately related to a mediating class B in the two premises, so reasoning is required to connect them. Many of the possible syllogisms that can be produced by varying the quantifier in each premise and the order of the terms actually yield no valid conclusion. One common finding is that subjects tend to endorse many conclusions as valid which do not necessarily follow from

their premises. However, error patterns are significantly related to the structure of the syllogisms presented. An old debate in the field concerns a dispute between the view that errors reflect an atmosphere bias (Woodworth & Sells, 1935) and the theory that they result from elicit conversion of premises (Chapman & Chapman, 1959). According to the atmosphere theory, subjects prefer conclusions which have a similar *mood* to their premises; for example, affirmative premises suggest affirmative conclusions. According to conversion theory, subjects tend wrongly to assume that *all* premises imply their converse statements, whereas some do not (All A and B, some A are not B). The two theories make notoriously similar predictions about problem difficulty and are thus hard to separate.

The details of these theories do not concern us here. We mention them because they have inspired the only work on instructional effects in syllogistic reasoning that we are aware of in the literature. Two studies compared the effects of "anti-atmosphere" and "anti-conversion instructions" but with contradictory findings. Simpson and Johnson (1966) claimed that anti-atmosphere instructions were effective in reducing atmosphere errors, but that the effects of anti-conversion instruction were unclear. In contrast, Dickstein (1975) reported that anti-conversion instructions were more effective in reducing errors in syllogistic inference. Perhaps the only safe conclusion we can draw from these two studies is that verbal instructions can significantly affect syllogistic reasoning.

Mediating between one theory and another of syllogistic reasoning is a rather narrow objective. In this study, we have chosen a phenomenon of much wider potential interest—the belief bias effect—which just happens to be studied normally with categorical syllogisms. Belief bias consists of a tendency to judge the argument of a syllogism as valid or invalid on the basis of whether or not it is *a priori* believable. Some illustrative data from the study of Evans, Barston and Pollard (1983) are shown in Table 1. Evans et al. presented their subjects with whole syllogisms—two premises and a single conclusion—and asked them to say whether or not the conclusion was valid, i.e. necessarily followed from the premises. An example of an invalid-believable problem used in this study, and frequently judged by subjects to be valid, is as follows:

No additive things are inexpensive
Some cigarettes are inexpensive
Therefore, some additive things are not cigarettes

Inspection of Table 1 reveals three important trends. First, we see that subjects accept far more valid than invalid conclusions, regardless of believability. This indicates that the subjects possess deductive competence and that a substantial amount of analytic reasoning is in fact taking place. Second, we note a massive belief bias effect evidenced by the much greater acceptance rate of believable than unbelievable conclusions. This occurs

TABLE 1
Percentage of Conclusions Judged to be Valid as a Function of Logical Validity and Believability in the Study of Evans et al. (1983): Combined Data from Three Experiments (n = 120)

	Believable	*Unbelievable*
Valid	89	56
Invalid	71	10

despite (1) the use of instructions which tell subjects to assume that the information given is true and to base their conclusions only on the premises, and (2) the use of a population group (undergraduate students) of well above average intelligence. Finally, we note also that there is an interaction between the two factors: the belief bias effect is more marked on invalid problems. A theoretical interpretation of these findings is offered below.

The belief bias effect in syllogistic reasoning is a suitable choice for the investigation of instructional effects for several reasons. First, the phenomenon is of general interest, since it concerns the idea that in reasoning from the information given we are unable to ignore our prior beliefs. It is important to know whether people can be persuaded to suspend their beliefs or whether the influence of knowledge on everyday reasoning is so pervasive as to be beyond conscious control. Second, unlike the Wason selection task, there is clear evidence that people can and do reason with some success on syllogistic tasks. This is clear from the substantial influence of validity on acceptance of conclusions shown in Table 2. Finally, the effects of instructions may shed some light on the debate between alternative theoretical accounts of the belief bias effect. At the very least, the manipulation should produce some new findings for the theorists to explain.

There are three main theories of the belief bias effect offered in the recent literature: the *selective scrutiny* and *misinterpreted necessity* models proposed in essence by Evans et al. (1983), but later named as such by Barston (1986) and Evans (1989), and the *mental models* theory (Oakhill & Johnson-Laird, 1985; Oakhill, Johnson-Laird, & Garnham, 1989). Recent discussions have indicated that no one of these theories can account for all of the findings in the literature (see Evans et al., 1993a, ch. 8; Newstead et al., 1992; Newstead & Evans, 1993; Oakhill & Garnham, 1993). However, the mental models theory can explain a majority of results and also incorporates aspects of both the other models. It also has the considerable advantage of providing a specific account of how syllogistic reasoning is performed (see Johnson-Laird & Bara, 1984). We there-

fore adopt a general mental models approach in this paper to provide a theoretical framework in which to understand our findings.

The mental models theory proposes that reasoning occurs in the following three main stages (see Johnson-Laird & Byrne, 1991):

- Reasoners construct a mental model to represent a possible state of the world in which the premises of the argument are true.
- Next, they form a putative conclusion which is true in the model and semantically informative, i.e. not a trivial inference such as the repetition of a premise.
- Finally, they ensure the deductive validity of the argument by attempting a search for counter-examples, i.e. models which are compatible with the premises and in which the putative conclusion does not hold. If no such counter-examples are found, then the inference is made.

Now let us return to the basic data of belief bias (Table 1). The model theory can account for the high acceptance of invalid–believable arguments on the grounds that when a putative conclusion is formed which is believable, the subject lacks motivation to search for counter-examples. In the study of Evans et al. (1983), invalid arguments always had a conclusion which *could* be true, given the premises, but need not be (i.e. counter-example models could be found). According to model theory, when the conclusions in such cases are unbelievable, subjects will look for the counter-examples and find them, hence rejecting invalid–unbelievable arguments. However, they will often accept believable conclusions uncritically (this is similar to the selective scrutiny of Evans et al., 1983). On valid arguments, however, counter-examples do not exist, so the effects of belief should be much less. This explains the belief × logic interaction observed by Evans et al.

While this argument is plausible, the model theory is less well equipped to explain why a significant (albeit weaker) belief bias should be observed on *valid* arguments, as was the case in the study of Evans et al. (1983). Oakhill et al. (1989) similarly found belief bias effects on simple one-model valid arguments, and proposed that unbelievable conclusions were removed by a post-reasoning "conclusion filter". This is a rare example of mental models theorists being forced to postulate a pure response bias mechanism extrinsic to the theory. However, it should also be said that belief bias effects are generally much weaker on valid arguments, and this finding of Evans et al. (1983) has not been consistently replicated (see, e.g. Newstead et al., 1992). We return to this issue below.

In the experiments reported here, we focus on the main source of belief bias—which is accounted for in the model theory—that is, the high rate of acceptance of invalid–believable arguments. The experimental manipulation is that of verbal instructions, which emphasises the concept of

logical necessity; that is, that conclusions should only be accepted if they *must* be true given the premises. It is reasonable to suppose that such instructional emphasis will make subjects more likely to search for counter-examples even when the conclusion is believable, thus reducing acceptance of invalid–believable arguments. Our starting point for this investigation is the final experiment of Newstead et al. (1992), who conducted an exploratory study of this kind following four other experiments which were generally supportive of the mental models account. In this experiment, a control group was presented with syllogisms using standard instructions, whereas an experimental group had instructions augmented by extra sentences designed to enhance understanding of logical necessity. For example, one elaboration, given only to the augmented instruction group, was:

> Please note that according to the rules of deductive reasoning, you can only endorse a conclusion which definitely follows from the information given. A conclusion that is merely possible, but not necessitated by the premises is not acceptable. Thus, if you judge that the information given is insufficient and you are not absolutely sure that the conclusion follows you must reject it and answer "NO".

Newstead et al. (1992) introduced this manipulation for theoretical reasons, arguing that the instructional emphasis on logical necessity should be effective according to two theories of belief bias—mental models and the misinterpreted necessity model—but not according to the selective scrutiny model. The above instruction clearly orients subjects towards the possibility that a conclusion which appears to follow at first sight, need not necessarily do so.

The results of Newstead and co-workers' experiment are summarised in Table 2. It can be seen that the data of the two groups differ noticeably

TABLE 2
Percentage of Subjects Accepting the Conclusion (i.e. Deeming it to be Valid), Shown for Each Problem Type, Divided According to Instruction Group (n = 24 in Each Instruction Group)[a]

	Believable	*Unbelievable*	*Neutral*
Standard instructions			
Valid	75	75	83
Invalid	50	0	29
Augmented instructions			
Valid	71	79	88
Invalid	17	4	25

[a]Data from Newstead et al. (1992, experiment 5).

in only one respect: there is a substantial drop in the frequency of acceptance of invalid–believable conclusions. While there was both a significant belief bias effect and belief × logic interaction under standard instructions, neither effect was significant under augmented instructions. There was no indication of a belief bias on valid arguments in either condition. Thus, these findings are strongly in line with the predictions of the mental models account. They also appear to contradict the general assertion of Evans (1989) that reasoning biases reflect pre-conscious heuristics which cannot be modified by verbal instruction. From a more general perspective, it appears that belief bias can be reduced or even eliminated by the use of verbal instruction.

It is important as a first step to attempt to replicate the effects of instructions reported by Newstead et al. and shown in Table 2. In designing the first experiment to be reported here, we were also conscious of the discrepancy in the findings of Newstead et al. (1992) and those of Evans et al. (1983) with regard to belief bias on valid arguments (compare Tables 1 and 2). The reasons for the discrepancy are not immediately apparent. The logical forms used by Newstead et al. were the same as in the earlier research (see Table 3) and the instructions for the control group were also the same as used by Evans et al. The main difference was that Newstead et al. used a new set of problem materials, introduced in order to add belief-neutral conclusions not employed by Evans et al. (1983). In both studies, believability of materials was determined by having statements rated by an independent group of subjects drawn from the same population as those taking part in the reasoning experiments. We will refer to the materials of Newstead et al. as the "new" materials and to those of Evans

TABLE 3
Logical Forms Used in the Experiments

Valid forms	
S1	No A are B Some C are B Therefore, some C are not A
S2	Some A are B No C are B Therefore, some A are not C
Invalid forms	
S3	Some A are B No C are B Therefore, some C are not A
S4	No A are B Some C are B Therefore, some A are not C

et al. as the "original" materials. Belief ratings for the original materials are presented by Evans et al. (1983).

Experiment 1 was therefore designed to provide a replication of experiment 5 of Newstead et al. (1992) with both new and original problem content. The belief-neutral condition was omitted, since this was not available in the original problem content.

EXPERIMENT 1

Method

Subjects. Sixty-four undergraduate psychology students at the University of Plymouth took part in partial fulfilment of a course credit requirement. They had no previous experience of this task, nor any training in logic.

Materials. The logical structures of the syllogisms used are shown in Table 3. These are similar to those employed by Evans et al. (1983) and have the advantage that the logic is not affected if subjects convert the premises (No A are B entails No B are A and so on), and that the same figures (i.e. order of the terms A, B and C) are used in both valid and invalid syllogisms. Half the subjects received syllogisms of the form S1 and S3, which involved C–A conclusions, and half received the S2 and S4 forms. Both original problem content (as used by Evans et al., 1983) and new problem content (as used by Newstead et al., 1992) were used.

An example of a syllogism used which was both valid (S1) and unbelievable, in new problem content, is as follows:

No animals are inhabitants of the island
Some tigers are inhabitants of the island
Therefore, some tigers are not animals

Design. The subjects were divided into two groups according to whether they received standard or augmented instructions. Each subject was asked to evaluate eight syllogisms, four using original and four using new content. The problems were presented in a booklet which contained two blocks of four problems. Half of each group received booklets with a block of original followed by a block of new problem content; the remaining half of the group received the reverse ordering of blocks. Within each block of problem content, the subjects received each of the four problem types: valid–believable, valid–unbelievable, invalid–believable and invalid –unbelievable. Combination of problem type and content was counterbalanced and presentation order was independently randomised within blocks for each subject.

Procedure. The subjects were run in groups of four. Each subject was given unlimited time to evaluate the validity of eight syllogisms. The standard instructions were based on those used by Evans et al. (1983, experiment 3, with references to the prose passage and verbalisation removed) and identical to those used by Newstead et al. (1992, experiment 5). The augmented instructions simply contained an additional passage outlining the principle of logical necessity with a short reminder at the end. These instructions were designed to follow a specific section of the instructions given by Dickstein (1981) and supplied by personal communication as they were not published in his paper.

The instructions were as follows. Sections in square brackets were given to the augmented instruction group only.

> This experiment is designed to find out how people solve logical problems. In the booklet which you have been given there are six logical reasoning problems. Your task is to decide whether the conclusion given below each problem follows logically from the information given in that problem.
>
> You must assume that all the information which you are given is true; this is very important. If, and only if, you judge that a given conclusion logically follows from the information given you should write "YES" in the space below the conclusion on that page. If you think that the given conclusion does not necessarily follow from the information given you should write "NO".
>
> [Please note that according to the rules of deductive reasoning, you can only endorse a conclusion if it definitely follows from the information given. A conclusion that is merely possible, but not necessitated by the premises is not acceptable. Thus, if you judge that the information given is insufficient and you are not absolutely sure that the conclusion follows you must reject it and answer "NO".]
>
> Please take your time and be certain that you have the logically correct answer before stating it.
>
> If you have any questions, please ask them now as the experimenter cannot answer any questions once you have begun the experiment.
>
> Please keep these instructions in front of you in case you need to refer to them later on.
>
> [REMEMBER, IF AND ONLY IF YOU JUDGE THAT A GIVEN CONCLUSION LOGICALLY FOLLOWS FROM THE INFORMATION GIVEN YOU SHOULD ANSWER "YES", OTHERWISE "NO".]
>
> Please do not turn back and forth from one problem to another once you have started. You must not make notes or draw diagrams of any kind to help you in this task.
>
> Thank you very much for participating.

Results

Table 4 sets out the percentage frequency of conclusions accepted, divided according to instructions and problem content and for combined content. The first observation is that response patterns across contents appear to be very similar.

In order to compare performance on controversial problems between instruction groups, three Mann–Whitney tests (two-tailed) were performed on logic, belief and interaction indices, calculated for each subject. The indices were calculated in the following manner:

- *Logic index*: total number of acceptances for invalid problems subtracted from that for valid problems.
- *Belief index*: total number of acceptances for unbelievable problems subtracted from that for believable problems.
- *Interaction index*: Total number of acceptances for the valid–believable problem, plus that for the invalid–unbelievable problem, subtracted from the total number of acceptances for the valid–unbelievable plus invalid–believable problems.

Comparing both types of content for each instruction group, no significant differences were found for logic, belief or interaction indices. It therefore seems that the original and new problem contents did not have

TABLE 4
Percentage of Conclusions Accepted for the Four Problem Types, Divided According to Content (Original *vs* New) and Instructions (Standard or Augmented) in Experiment 1 (n = 32 in Each Instruction Group)

	Standard		*Augmented*	
	Believable	*Unbelievable*	*Believable*	*Unbelievable*
Original content[a]				
Valid	84	78	78	84
Invalid	63	13	34	0
New content				
Valid	84	81	84	84
Invalid	47	19	53	9
Combined content				
Valid	84	80	81	84
Invalid	55	16	44	5

[a]As used in the study of Evans et al. (1983).

differential effects on overall acceptance rates for the four problem types. However, a within-subject design was used and overall response patterns could have been distorted by carry-over effects. Since all subjects received a block of both problem contents, the two content blocks in each booklet were separated and compared, but the results were very similar in each condition and to the combined figures shown in Table 4. Consequently, all further analyses were carried out on data for both original and new contents combined.

In contrast with the findings of Newstead et al., sign tests revealed significant effects of logic, belief and a logic × belief interaction for *both* instruction groups ($P < 0.001$ in all cases, except for the belief × logic interaction in the standard instruction group, where $P < 0.01$; all based on one-tailed tests). Mann–Whitney tests on logic, belief and interaction indices comparing standard and augmented instruction groups also revealed no significant differences. Hence, Newstead and co-workers' finding of a reduced belief index under augmented instructions was not replicated.

Discussion

The objectives of Experiment 1 were (1) to replicate the finding of Newstead et al. (1992, experiment 5) that the belief × logic interaction could be reduced by use of augmented instructions emphasising logical necessity, and (2) to examine whether possible discrepancies with the findings of Evans et al. (1983) were due to the change of problem materials. With respect to the second objective, the results are quite clear. The behaviour observed with both the new and original materials was highly similar throughout. Moreover, the present findings conform with those of Newstead et al. (and conflict with those of Evans et al.) in finding that there is no apparent belief bias effect for valid syllogisms. It is curious that this aspect of the findings of Evans et al. (1983)—which they found across three separate experiments—should not replicate even with a virtually identical repetition of their conditions of testing. However, the absence of such a finding is more theoretically convenient for the mental models account, as we showed earlier.

The ineffectiveness of the augmented instructions in Experiment 1, however, could be seen as evidence for the view that biases *are* caused by pre-conscious heuristics which are resistant to instructional manipulations, as proposed by Evans (1989). The finding is compatible with the mental models theory only if one supposes that the mechanism of searching for counter-examples is entirely unconscious and hence inaccessible to verbal instructions. The relation of mental models reasoning to consciousness is unclear in the theoretical accounts offered by Johnson-Laird (1983;

Johnson-Laird & Byrne, 1991). While he argues that models may have a conscious representation in the form of an image, he also says they need not be conscious at all. At the same time, it is claimed that models occupy space in working memory, which is generally viewed by cognitive theorists as equating with conscious access (see Ericsson & Simon, 1980).

However, as we see from Table 2, the instructional manipulation did reduce belief bias in experiment 5 of Newstead et al. (1992), so we have a discrepancy between the findings of their experiment and ours to explain. One possibility that we have eliminated is that the new problem materials have a different character from the original ones of Evans et al. This leaves us with two further possibilities. Either the presence of neutral conclusions —not included in Experiment 1—affected responding, or else the discrepancy in findings was due to statistical error. Hence, Experiment 2 repeats the crucial augmented instructions condition of Experiment 1 using new problem content and with belief-neutral conclusions included in order to constitute an exact replication of the experiment reported by Newstead et al. (1992). In an attempt to avoid problems of insufficient statistical power, a substantially larger sample size was used than that employed by Newstead et al. (72 compared with their 24). The main objective was to see whether the belief and belief × logic interaction effects could be demonstrated despite augmented instructions, as was the case in Experiment 1. A secondary objective was to provide further investigation of reasoning with belief-neutral conclusions.

EXPERIMENT 2

Method

The materials and syllogisms used were identical to those of Newstead et al. (1992, experiment 5). They were the same as those of the augmented instructions group in Experiment 1 using new materials only, and with the addition of problems with belief-neutral conclusions. The experimental design also differed from that of Experiment 1 in that no standard instruction group was included. In total, 72 undergraduate psychology students at the University of Plymouth served as subjects. They had no previous experience of this task, nor any training in logic.

Results and Discussion

The pattern of responses over the six different problems is set out in Table 5. Disregarding the belief-neutral conclusions for the moment, the data were analysed for effects of logic, belief and a logic × belief interaction using one-tailed sign tests. The crucial comparison is with the findings of

TABLE 5
Percentage of Subjects Accepting the Conclusion (i.e. Deeming it to be Valid) in Experiment 2 (n = 72)

	Believable	*Unbelievable*	*Neutral*
Valid	72	64	75
Invalid	35	8	43

Newstead et al. under augmented instructions (see Table 2), as they performed similar analyses. As in their study, we found a significant effect of logic on responding ($P < 0.001$, sign test). However, Newstead et al. found neither a significant effect of belief as a main effect, nor a significant interaction of belief and logic in this condition. By contrast, in Experiment 2, there was a highly significant effect of belief ($P < 0.001$) and the belief × logic interaction approached significance ($P = 0.008$).

In conjunction with the findings of Experiment 1, we are forced to conclude that the lack of belief bias under augmented instructions reported by Newstead et al.—based on a sample of 24 subjects—is a type 2 statistical error. Note that the rate of acceptance of invalid–believable conclusions in their study (17%) was higher than for invalid–unbelievable conclusions (4%), although this was not sufficient to produce any significant differences. In Experiments 1 and 2 of this paper, however, we found that (1) significant effects of belief bias were maintained under augmented instructions (Experiments 1 and 2), (2) no significant reduction of the belief × logic interaction occurred when compared with reasoning under standard instructions (Experiment 1), and (3) there was a significant (Experiment 1) and near significant (Experiment 2) belief × logic interaction under augmented instructions.

The acceptance rates for belief-neutral conclusions (see Table 5) broadly replicate the findings of Newstead et al. In both studies, considerably more subjects endorsed such conclusions on valid than on invalid arguments. However, the interesting comparison is with the rates for acceptance on logically equivalent believable and unbelievable conclusions. Far from being intermediate in endorsements, both Experiment 2 and the corresponding augmented instructions condition of Newstead et al. (see Table 2) show that neutral conclusions are accepted (non-significantly) more often than are believable ones. The conclusion suggested is that augmented instructions are not sufficient to suppress endorsement of fallacious conclusions unless those conclusions are *a priori* unbelievable (Oakhill & Garnham, 1993, make a similar point). In terms of mental models theory, this implies that search for counter-examples is not an habitual strategy of

reasoning, but rather one that requires motivating. We return to this point in the General Discussion.

EXPERIMENT 3

Having established that augmented instructions do not eliminate effects of belief bias, as suggested by the findings of Newstead et al., we now consider the weaker claim that these effects are *reduced* by the use of augmented instructions. This is important, since some debiasing is clearly better than none. Evidence in favour of the hypothesis is provided by the data of Newstead et al. (see Table 2), further analysis of which has revealed that the belief index was significantly lower in the augmented instruction group. However, no differences were found in Experiment 1 of this paper (see Table 4). Comparison of Table 5 with Table 1, however, reveals a much lower acceptance of invalid–believable conclusions than in the study of Evans et al. (1983). This supports the hypothesis that belief bias—especially on invalid arguments—may be reduced by augmented instructions, although such comparisons between different experiments are hazardous.

Let us assume for the moment that there is a real, if weak, tendency for the acceptance of invalid–believable conclusions to be reduced by augmented instructions. It may be, as we have supposed, that this was due to the instructional emphasis placed upon the concept of necessity in these instructions. Alternatively, the effect of elaborating the instructions might be a more general one of inducing the subject to adopt a more cautious attitude. Such a finding would be of more interest with regard to the general issue of whether instructions can debias reasoning, suggesting a greater generality of results.

Two of the experiments reported by Barston (1986, experiments 4 and 5) are pertinent to the argument here. These experiments tested for belief bias in a syllogistic production task, but unlike other studies (Markovits & Nantel, 1989; Oakhill & Johnson-Laird, 1985) failed to find any. One possible reason for the lack of belief bias was the fact that Barston used very complex instructions, which strongly emphasised the structure of the syllogism, explained the logical meaning of "SOME" and repeatedly emphasised the need to base responses only on the information given. These instructions did not, however, give the specific additional emphasis to the concept of logical necessity as in the augmented instructions groups of the preceding experiments.

Hence Experiment 3 compares performance under the "complex" instructions of Barston (1986) with the "simple" instructions which were a reduced and simplified form of these. If complex instructions induce a more cautious approach, then we would expect a drop in acceptance of

invalid–believable conclusions with a consequent reduction in both belief bias and the belief × logic interaction.

Method

Subjects. Thirty-two undergraduate psychology students at the University of Plymouth acted as paid volunteers. They had no previous experience of this task, nor any training in logic.

Design. All subjects received four syllogisms to evaluate, which consisted of the four problem types: valid–believable, valid–unbelievable, invalid–believable and invalid–unbelievable. All problems were in "figure 3" format. The problem content was identical to that used by Evans et al. (1983). Each problem type occurred equally with each problem content, and presentation order was randomised. The problems were presented in booklet form and a space was provided beneath each conclusion for responses ("Yes" or "No") to be written.

Procedure. The subjects were run in groups of four and divided into two groups. One group received simple instructions, and the other received complex instructions. The complete instructions were as follows:

Simple instructions

This experiment is designed to find out how people solve logical problems. You will be tested on four logical reasoning problems, which are contained within the booklet which you have been given. Your task is to decide whether or not a given conclusion follows logically from the information given—and this information only. You must assume that all the statements within the problem are true—this is very important. *If, and only if, you judge that the given conclusion logically follows from the statements given you should answer by writing "YES" below the conclusion, otherwise write "NO".*

Please take your time and be certain that you have the logically correct answer before stating it.

If you have any questions, please ask them now, as the experimenter cannot answer any after you have started.

Please keep these instructions in front of you in case you need to refer to them later on.

REMEMBER, IF AND ONLY IF YOU JUDGE THAT THE GIVEN CONCLUSION LOGICALLY FOLLOWS FROM THE STATEMENTS GIVEN YOU SHOULD ANSWER "YES", OTHERWISE "NO".

Please do not turn back and forth from one problem to another once you have started. You must not make notes or draw diagrams of any kind to aid you in this task.

Complex instructions

This experiment is designed to find out how people solve logical problems. In the booklet which you have been given there are 4 logical reasoning problems. Your task is to decide whether or not the conclusion given does or does not logically follow from the information which is given above. The information takes the form of two statements (premises) which can be expressed symbolically as follows:

ALL B ARE A,
SOME C ARE B.

As you can see, the two premises tell us something about the relationship between three terms: A, B and C. The term B never appears in the conclusion, since the conclusion is a statement about the relationship between A and C, or vice versa. The conclusion to the above example is, therefore, "SOME A ARE C".

Since this is a problem requiring logical analysis, you should interpret the word "SOME" in its strictly logical sense; meaning *AT LEAST ONE AND POSSIBLY ALL*. So the statement "SOME B ARE C" does not necessarily also mean that "SOME B ARE NOT C".

In the booklet you will find four different logical problems. They are the same type of problem as the example problem which is shown above; however, the terms used will not be letters of the alphabet, but real words instead. Your task is to write down, below the conclusion given, "YES" if you judge that the conclusion necessarily follows from the information given, or "NO" if you judge that the conclusion does not necessarily follow from the information given.

You are reminded that you must base your decision on the information given in the two premises—and this information only. You must assume that all the information which you are given is true—this is very important. If, and only if, you judge that a specific conclusion logically follows from the information given you should write "YES"; the conclusion given may not always be the correct one.

Please take your time and be certain that you have made the logically correct decision before stating it.

If you have any questions, please ask them now, as the experimenter cannot answer any questions once you have begun the experiment.

Please keep these instructions in front of you in case you need to refer to them later on.

REMEMBER, YOUR DECISION SHOULD BE BASED SOLELY UPON WHAT CAN BE DEDUCED WITH ABSOLUTE CERTAINTY FROM THE TWO PREMISES—AND THIS INFORMATION ONLY.

Please do not turn back and forth from one problem to another once you have started. You must not make notes or draw diagrams of any kind to aid you in this task.

Results and Discussion

The frequency of acceptance of conclusions for both instruction groups is shown in Table 6. Sign tests revealed an effect of logic under both simple and complex instructions ($P < 0.05$, one-tailed tests for both groups). An effect of belief was observed under simple instructions ($P < 0.05$), although for complex instructions the effect fell just short of significance ($P = 0.055$, one-tailed test). An interaction between the two effects was present under simple instructions ($P < 0.01$), but no significant interaction was observed under complex instructions.

Comparisons between instruction groups revealed no significant differences for logic indices. This was also true for belief indices. However, there was a significant difference for interaction indices ($P < 0.05$). (All comparisons are based on one-tailed Mann–Whitney tests.)

TABLE 6
Percentage of Subjects Accepting the Conclusion for Each of the Four Problem Types, Divided According to Instruction Group: Experiment 3 ($n = 32$ in Each Instruction Group)

	Simple Instructions		*Complex Instructions*	
	Believable	*Unbelievable*	*Believable*	*Unbelievable*
Valid	81	69	81	63
Invalid	81	19	44	31

Although there were no significant differences on belief indices, belief bias fell short of significance under complex instructions and the belief × logic interaction index was also significantly lower for this group. Inspection of Table 6 reveals that the principal difference between the two groups was, again, in the acceptance rate of invalid–believable conclusions which was considerably lower in the complex instruction group. Note also, that for the first time in the present study, a (weak) trend towards a belief bias on valid arguments appears in the data.

GENERAL DISCUSSION

The principal objective of this study was to discover whether it is possible to reduce or eliminate reasoning biases by use of verbal instruction. The belief bias effect was chosen for study because it has been shown to be very robust and pervasive and because it clearly co-occurs with logical reasoning processes which lead subjects to correct solutions part of the

time. The three experiments reported here taken in conjunction with experiment 5 of Newstead et al. (1992) provide a clear answer: the bias may be reduced, but not eliminated by instructional manipulation. In particular, the acceptance of invalid–believable arguments is reduced, though not sufficiently (statistical power permitting) to eliminate evidence of some belief bias effect. It also appears, from Experiment 3, that the reduction is achieved by elaboration of logical principles in general and does not require a specific emphasis on the principle of logical necessity.

It is interesting that instructions can reduce the bias, but also interesting that they cannot eliminate it. Taking the second aspect first, it provides further evidence of how remarkably robust and powerful is the belief bias effect in syllogistic reasoning. The bias has once again been demonstrated across three experiments in which intelligent adults are clearly instructed to perform deductive reasoning. Even the simple instructions of Experiment 3 contained the passage, "You must assume that all the information you are given is true; this is very important. If, and only if, you judge that the conclusion follows logically from the given information you should write 'YES' below the conclusions, otherwise write 'NO' ", and the standard instructions of the other experiments contained a similar passage. Significant belief bias effects were also found in Experiments 1 and 2, despite the use of augmented instructions, and the effect was marginally significant under the complex instructions of Experiment 3. The first conclusion from our study, then, must be that subjects find it extraordinarily difficult to suspend their beliefs when attempting a reasoning task, even when exhorted to do so. We consider a possible explanation for this below.

The fact that instructions can reduce the bias—as was clearly shown by both Newstead et al. (1992, experiment 5) and our own Experiment 3—is, of course, encouraging to those interested in debiasing. Our findings are more compatible with the optimistic results of Fong et al. (1986) on statistical reasoning, than with the pessimistic findings of Cheng et al. (1986) on deductive reasoning. As stated earlier, their choice of the Wason selection task—which provides no undisputed evidence of logical reasoning—was indecisive with regard to the question of whether logical reasoning can be debiased by instruction. We now turn to a consideration of the theoretical implications of our findings.

In this paper, we have adopted the mental models theory as the general framework within which to view and interpret our findings. The results bear in particular on two issues pertinent to this theory: (1) the postulated search for counter-examples and (2) the degree of conscious control of the reasoning process. With regard to the first issue, the finding that belief bias is maintained with neutral conclusions suggests that subjects will accept putative conclusions without an attempt to verify them, *unless* they are

motivated by an unbelievable conclusion. This is rather different from the original proposal that conclusions are accepted unverified because they are believable. It suggests that subjects do not habitually search for counter-examples and are hence less deductively competent than originally proposed. This conclusion receives support from the recent discussion of models theory and rationality proposed by Johnson-Laird and Byrne (1993), in which they now appear to argue that while inferences are suppressed by the availability of counter-examples, subjects are *not* generally good at seeking such cases out. In addition to accounting for the very common endorsement of fallacious inferences in both syllogistic and conditional reasoning experiments (see Evans et al., 1993a), this modification to the theory allows for an account of inductive inference based on much the same method of reasoning (see Johnson-Laird, 1993).

Our findings do, however, indicate that in addition to the presentation of unbelievable conclusions, the search for counter-examples may be facilitated by elaborated verbal instructions; hence the reduction in acceptance of invalid–believable conclusions observed in some of our experiments. Thus it seems that while subjects *can* seek to refute a putative conclusion, it is quite difficult to persuade them to do so. The fact that instructions can be effective suggests at least some degree of conscious control of the reasoning process—an unclear area, as we have said earlier, in the proposals of mental models theorists. Belief bias cannot be entirely attributable to the influence of an unconscious heuristic as suggested by Evans (1989). However, as we have also shown that the bias is not eliminated by such instructions, we must ask why it is so pervasive.

A likely account rests on the assumption that subjects bring to the laboratory habitual methods of reasoning which are effective in the real world and which they find very hard to alter in response to experimental instructions. Evans, Over and Manktelow (1993b) have argued that logicality is a poor criterion against which to assess the rationality of everyday reasoning, which depends instead on the effectiveness with which it leads an individual to achieve his or her goals. Hence, they argue, illogical reasoning observed in the laboratory may be far more rational than it appears. In a discussion of the belief bias effect, they suggest effectively that people reason in a broadly Bayesian manner. That is to say, new evidence is weighted against prior beliefs in assessing one's posterior belief in a proposition. Although experimental subjects are clearly instructed to ignore prior beliefs and base their reasoning on the information presented, it is possible that they persist with this kind of approach. If the argument presented is perceived as "evidence", then an invalid argument might be regarded as providing no useful information. In this case, subjects may be expected to retain their prior belief, thus accounting for the particularly marked influence of belief on invalid arguments. According to this view,

the deductive reasoning that the subject engages in is only one input to the *decision* they make as to whether or not to accept the conclusion. Thus in so far as our instructions are effective, this may lie in persuading subjects to accept the experimental task and base their conclusions on the immediate reasoning task in front of them. We are reminded here of the suggestion in Henle's (1962) classic paper that subjects may "fail to accept the logical task".

A potential difficulty for a universal reasoning theory such as that of mental models lies in accounting for individual differences in performance (see Roberts, 1993). In the current experiments, as in others in the literature, the subjects produce a variety of responses to problems and the effects of our manipulations are seen in the form of statistical trends. For example, some subjects succeed in rejecting invalid conclusions even under conditions where they are very common. In general, model theorists suggest that individuals may vary in how competently or completely they execute the stages of reasoning proposed by the theory. However, Evans et al. (1983) have shown strong evidence that the conflict between belief and logic in belief bias experiments occurs *within* individual subjects. First, they showed through verbal protocol analysis that, *on a given problem*, subjects who refer to the premises of an argument rather than just its conclusion are less likely to show belief bias. Surprisingly, however, they could not find evidence for consistent individual differences in reasoning strategy.

In one striking analysis, Evans et al. (1983, p. 303) showed that the probabilities of accepting conclusions of different syllogisms were statistically independent. In other words, knowing that a subject shows belief bias on one problem does not enable us to predict whether they will do this on another. There is nothing intrinsic to the mental models theory which helps us to understand this. However, this within-subject conflict *is* compatible with the account offered above, in which we suggest that individual subjects are influenced both by an attempt to reason as instructed and by a habitual tendency to incorporate prior beliefs into the assessment of new evidence. The effect of elaborated instructions is seen in this context as shifting the balance—though rarely decisively—in favour of reasoning. The mental models theory is seen as providing only a partial account of what is going on—the attempt to reason.

What implications do our findings have for the teaching of thinking skills? Clearly, there is a difficulty in persuading people to think logically about a new problem when they have preconceived beliefs about it. Detailed instructional emphasis helps but may not be sufficient. One way we can exploit the belief bias effect, however, is to use the finding that reasoning is most effective when people are motivated by an unbelievable conclusion. Under these conditions, fallacious arguments are usually

rejected, while there is only a weak tendency to reject valid arguments. Hence, people may better develop critical reasoning when confronted with arguments which are counter-factual or contrary to their prior beliefs. For example, if you want to develop critical thinking about research design methodologies, you should ask your students to read studies whose conclusions they disagree with. Specific evidence that this may affect critical thinking is provided by the study of Lord, Ross and Lepper (1979). Hence, when you want people to be critical, the belief bias effect may be turned to positive advantage.

In conclusion, there is a presumption among educationalists and psychologists interested in the teaching of thinking that verbal instructions in principles of reasoning or decision making is beneficial. Much research of this kind is done in classroom settings, which has the advantage of high external validity. However, the research reported in this paper brings a new perspective to the topic by investigating the effectiveness of instructions in the face of a powerful and well-established cognitive bias. Belief bias and closely related phenomena such as confirmation bias must presumably influence thinking in many real-world domains. Our findings suggest that such biases are very hard to eradicate and that verbal instruction in relevant principles may exert only a moderate restraining influence.

Manuscript received October 1993
Revised manuscript received February 1994

REFERENCES

Baron, J. (1985). *Rationality and intelligence*. Cambridge: Cambridge University Press.

Baron, J. (1988). *Thinking and deciding*. Cambridge: Cambridge University Press.

Barston, J.I. (1986). *An investigation into belief biases in reasoning*. Unpublished PhD thesis, University of Plymouth.

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

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

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

Dickstein, L.S. (1981). Conversion and possibility in syllogistic reasoning. *Bulletin of the Psychonomic Society, 18*, 229–232.

Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as data. *Psychological Review, 87*, 215–251.

Evans, J.St.B.T. (1989). *Bias in human reasoning: Causes and consequences*. Hove: Lawrence Erlbaum Associates Ltd.

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

Evans, J.St.B.T., Newstead, S.E., & Byrne, R.M.J. (1993a). *Human reasoning: The psychology of deduction*. Hove: Lawrence Erlbaum Associates Ltd.

Evans, J.St.B.T., Over, D.E., & Manktelow, K.I. (1993b). Reasoning, decision making and rationality. *Cognition, 49*, 165–187.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds), *Judgement under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Fong, G.T., Krantz, D.H., & Nisbett, R.E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18*, 253–292.

Gorman, M.E., & Gorman, M.E. (1984). Comparison of disconfirmatory, confirmatory and control strategies on Wason's 2–4–6 task. *Quarterly Journal of Experimental Psychology, 36A*, 629–648.

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

Holland, J.H., Holyoak, K.J., Nisbett, R.E., & Thagard, P.R. (1986). *Induction: Processes of inference, learning and discovery*. Cambridge, MA: MIT Press.

Johnson-Laird, P.N. (1983). *Mental models*. Cambridge: Cambridge University Press.

Johnson-Laird, P.N. (1993). *Human and machine thinking*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Johnson-Laird, P.N., & Bara, B.G. (1984). Syllogistic inference. *Cognition, 16*, 1–62.

Johnson-Laird, P.N., & Byrne, R.M.J. (1991). *Deduction*. Hove: Lawrence Erlbaum Associates Ltd.

Johnson-Laird, P.N., & Byrne, R.M.J. (1993). Models and deductive rationality. In K.I. Manktelow & D.E. Over (Eds), *Rationality*. London: Routledge.

Lord, C., Ross, L., & Lepper, M.R. (1979). Biased assimilation and attitude polarisation: The effect of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*, 2098–2109.

Manktelow, K.I., & Over, D. (Eds) (1993). *Rationality*. London: Routledge.

Markovits, H., & Nantel, G. (1989). The belief bias effect in the production and evaluation of logical conclusions. *Memory and Cognition, 17*, 11–17.

Newstead, S.E., & Evans, J.St.B.T. (1993). Mental models as an explanation of belief bias effects in syllogistic reasoning. *Cognition, 46*, 93–97.

Newstead, S.E., Pollard, P., Evans, J.St.B.T., & Allen, J. (1992). The source of belief bias in syllogistic reasoning. *Cognition, 45*, 257–284.

Nickerson, R.S., Perkins, D.N., & Smith, E.E. (1985). *The teaching of thinking*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Oakhill, J., & Garnham, A. (1993). On theories of belief bias in syllogistic reasoning. *Cognition, 46*, 87–92.

Oakhill, J., & Johnson-Laird, P.N. (1985). The effect of belief on the spontaneous production of syllogistic conclusions. *Quarterly Journal of Experimental Psychology, 37A*, 553–570.

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

Roberts, M.J. (1993). Human reasoning: Deduction rules or mental models or both. *Quarterly Journal of Experimental Psychology, 46A*, 569–590.

Simpson, M.E., & Johnson, D.M. (1966). Atmosphere and conversion in syllogistic reasoning. *Journal of Experimental Psychology, 72*, 197–200.

Wason, P.C. (1966). Reasoning. In B.M. Foss (Ed.), *New horizons in psychology I*. Harmondsworth: Penguin.

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