THINKING & REASONING, 2004, *10* (1), 31–53

Does everyone love everyone? The psychology of iterative reasoning

Paolo Cherubini
Università di Milano-Bicocca, Italy

P. N. Johnson-Laird
Princeton University, NJ, USA

When a quantified premise such as: Everyone loves anyone who loves someone, occurs with a premise such as: Anne loves Beth, it follows immediately that everyone loves Anne. It also follows that Carol loves Diane, where these two individuals are in the domain of discourse. According to the theory of mental models, this inference requires the quantified premise to be used again to update a model of *specific* individuals. The paper reports four experiments examining such iterative inferences. Experiment 1 confirmed that they are harder than immediate inferences. Experiment 2 extended the finding to negative inferences, i.e., granted that Anne does not love Beth, it follows from the quantified premise that Carol does not love Diane. Experiment 3 established that intermediate steps referring to specific individuals are accepted more readily than intermediate steps referring to quantified variables. Experiment 4 showed that the participants' written justifications corroborated the model theory.

Correspondence should be addressed to Paolo Cherubini, Dipartimento di Psicologia, Università di Milano-Bicocca, 1, P.za dell'Ateneo Nuovo, 20126 Milano (Italy). Email: paolo.cherubini@unimib.it

This paper was made possible by the Department of Developmental and Social Psychology at the University of Padua, which provided a grant to the first author in September – December 2000, and which hosted the second author in October 2000. The research was also supported in part by a grant to the second author from the National Science Foundation (Grant BCS 0076287) to study strategies in reasoning. Special thanks are due to the Head of the Department, Professor Alberto Mazzocco, and to Professors Paolo and Maria Legrenzi for their hospitality. We also thank Ketti Mazzocco and Luigi Pellizzaro for their help in collecting the data, and Marcin Hitczenko for implementing the computer program of the model theory.

http://www.tandf.co.uk/journals/pp/13546783.html DOI: 10.1080/13546780342000025

Imagine a world in which there are four people: Anne, Beth, Carol, and Diane, and in which the following two assertions are true:

Everybody loves anyone who loves someone.
Anne loves Beth.
Does it follow that everyone loves Anne?

A moment's thought should convince you that indeed the conclusion is valid, i.e., given the truth of the premises, then the conclusion must be true. Now, consider the same world and premises but a different question:

Does it follow that Carol loves Diane?

Your immediate intuition is likely to be that the conclusion does not follow validly: whether or not Carol loves Diane seems quite independent of the information in the premises. But, in fact, the inference *is* valid. It seems harder to grasp its validity. Why? This is the question that the present paper addresses. Its answer improves our understanding of human reasoning.

What do people have to do in order to make the preceding inferences? Various psychology theories of reasoning have been based on formal derivations akin to those of logical proofs (e.g., Braine & O'Brien, 1998; Rips, 1994). Such theories postulate that the preceding inferences depend on formal derivations. We return to this possibility later in the paper, but we now consider an account based on the theory of mental models. According to this theory, naïve reasoners—those who have not mastered logic, make inferences by imagining the possibilities compatible with the premises (see, e.g., Johnson-Laird & Byrne, 1991). With the example above, the first step is to use the two premises to infer that everyone loves Anne (the *immediate* conclusion). It follows in an *intermediate* conclusion that Diane loves Anne. This intermediate conclusion does not require any further effort, because its model is embedded in the representation of the state of affairs in which everyone loves Anne. Next, an iterative use of the general premise yields the further *intermediate* conclusion that everyone loves Diane. This second intermediate conclusion requires some effort, because it depends on updating the initial model. The second *intermediate* conclusion yields the *iterative* conclusion: Carol loves Diane.

In general, the theory assumes that reasoners try to build as few models as possible and as simple models as possible (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). The premise that there are four people: Anne, Beth, Carol, and Diane, elicits a corresponding mental model:

(0) person Anne
person Beth
person Carol
person Diane

The *quantified* premise: Everybody loves anyone who loves someone, calls for the progressive construction of a model:

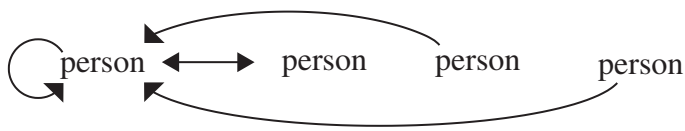
Everyone: each member of the set of persons
loves: is in the relation of *loving* to
anyone: each member of the set of persons
who: that is
loves: in the relation of *loving* to
someone: at least one member of the set of persons.

"Everyone" refers to a set of persons, that is Anne, Carol, Beth and Diane. "Anyone who loves someone" refers to a *second* set of persons, whose members share a property (*to love someone*). The second set could be empty. But, if there is at least one member of the second set, then everyone loves this person. This is an "antecedent" possibility, because it states a condition in which a consequent holds. The antecedent possibility is represented in the following model (in which the arrow denotes the relation of *loving*):

(1) person→person person person

where one person (on the left of the arrow) loves one other person (on the right of the arrow). There is an alternative possibility in which the relation is false, i.e., the set is empty because noone loves anyone. Any instance of the possibility (modelled in 1), however, satisfies the consequent possibility of the quantified premise, i.e. that a person who loves someone is loved by everyone:

(2)



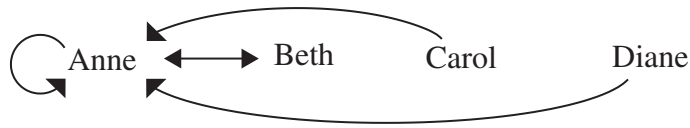
The relations asserted in the quantified premise do not depend on the specific individuals: each individual denoted by "person" can be Anne, Beth, Carol, or Diane, but there can be only one individual for each name, as in model 0. The *specific* premise that Anne loves Beth is represented by the model:

(3) Anne→Beth

However, since model 0 represents that Anne is a person and Beth is a person, model 3 directly matches the antecedent possibility of model 2. The

remaining persons in model 2, by exclusion, are matched to Carol and Diane. The resulting model is:

(4)



This initial model (4) of the two premises yields the *immediate* conclusion:

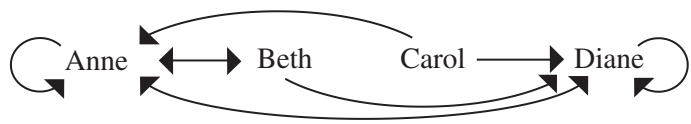
Everyone loves Anne.

The preceding model (4) also supports the intermediate conclusion:

Diane loves Anne.

This intermediate conclusion requires little further work to infer, because it is represented in model 4. However, Diane satisfies the antecedent possibility (1) of the quantified premise, and so provided that the reasoner notices this intermediate conclusion and remembers the original meaning of the quantified premises, the consequent possibility (2) of the quantified premise can be used in a second iteration to update the model of the premises:

(5)



This model yields the intermediate conclusion:

Everyone loves Diane

and the iterative conclusion follows at once:

Carol loves Diane.

By the same sort of argument, a further iterative use of the quantified premise on the preceding intermediate conclusion yields the conclusion that everyone loves everyone (including themselves), and at this point the model cannot be expanded any further by an iterative use of the quantified premise.

A computer program written by Marcin Hitczenko implements the essentials of the model theory's account of these inferences.

The second intermediate conclusion and the iterative conclusion can be drawn only if the initial model of the premises (4) is updated by an iterative use of the quantified premise. There are at least three reasons why this step is less likely to occur than the initial step. First, reasoners may operate on the tacit assumption that once a premise has been used in an inference, there is no need to use it again. Second, the iterative step depends on the existence of the initial model (4); the presence of this model in working memory imposes a load on the processing capacity of working memory. This load is likely to impede any further process of inference. In other words, after building the initial model, instantiated with specific individuals, naïve reasoners are likely to forget the original meaning of the quantified premise, and so they do not grasp that each of the intermediate conclusions from the initial model (such as, "Diane loves Anne") satisfies the antecedent possibility of the quantified premise. Third, individuals are less likely to expand models when they have already constructed a model that yields a conclusion (Cherubini, Garnham, Oakhill, & Morley, 1998; Oakhill, Garnham, & Johnson-Laird, 1990). In sum, the model theory makes two main predictions. First, the immediate conclusion should be easier to draw than the iterative conclusion. Second, the locus of difficulty should be the iterative updating of the initial model, that is in the inference of the intermediate conclusion (2) from the intermediate conclusion (1). There should be little residual difficulty in the other reasoning steps. The following experiments were designed to test these predictions, and Experiments 1 and 2 tested the first prediction.

EXPERIMENT 1

Method

Design. The participants acted as their own controls and carried out five trials with each of six sorts of inference. The six sorts of inference were based on pairing an immediate and an iterative conclusion on separate trials with three pairs of premises. The immediate conclusion could be drawn from an initial model of the premises but the iterative conclusion called for an iterated use of the quantified premise. The experiment was carried out in Italian from which we have here translated the six sorts of problem:

(i) Everyone *verbs* anyone who *verbs* someone.
A *verbs* B.
Does it follow that everyone *verbs* A? (Immediate conclusion)
Does it follow that C *verbs* D? (Iterative conclusion)

(ii) C *verbs* all those who *verb* D.
D *verbs* herself.
Does it follow that C *verbs* D? (Immediate conclusion)
Does it follow that C *verbs* C herself? (Iterative conclusion)

(iii) Anyone who is *verbed verbs* everyone.
A *verbs* B.
Does it follow that B *verbs* everybody? (Immediate conclusion)
Does it follow that C *verbs* D? (Iterative conclusion)

There were five versions of each of these problems created using the following verbs: *loves, admires, respects, esteems*, and *appreciates*. In order to have an equal number of invalid inferences, there were 30 filler inferences based on the same premises but combined with questions about invalid conclusions, for example:

Everyone *verbs* anyone who *verbs* someone.
A *verbs* B.
Does it follow that not everyone *verbs* A? (Immediate conclusion)
Does it follow that not everyone *verbs* D? (Iterative conclusion)

Although we have labelled the invalid conclusions as "immediate" and "iterative", their status differs from the corresponding valid conclusions: participants could respond "no" correctly to the invalid conclusions, either because they thought that the conclusion was possible but not necessary, or because they inferred that the conclusion was impossible. Hence, the difficulty of invalid and valid inferences is not comparable, and so the invalid inferences were fillers to balance the proportion of correct "yes" and "no" answers. The resulting 60 inferences were presented in a different random order to each of the participants.

Participants. A total of 20 students (mean age 21.7 years) of the University of Padua volunteered to participate in the experiment, which lasted for about half an hour. None had taken any course in logic or the psychology of reasoning.

Procedure. The experiment was carried out in Italian under the control of an Apple computer running the PsychLab program (Bub & Gum, 1991). The key instructions translated from the Italian were as follows:

In this experiment you will be presented with some premises. Please read them carefully, and take care to understand their meaning. The premises will remain on the screen for the full duration of each trial. When you are confident about the meaning of the premises, press the space bar on the keyboard: A conclusion will appear in the space below the premises. If the conclusion follows from the premises, press the "YES" key on the keyboard; otherwise, press the "NO" key on the keyboard.

The participants were also told to try to respond correctly, but as fast as possible.

Each trial began with the sentence, ''Imagine a world in which there are four persons''. There followed at 1-second intervals a list of four first names, either all feminine or all masculine, the quantified premise, and the specific premise, which remained on the screen until the end of the trial. After the participants had read and understood the premises, they pressed the space bar, which led to the presentation of the putative conclusion at the bottom of the screen: ''Does it follow that . . .?''

The participants responded by pressing either a key labelled ''YES'' or a key labelled ''NO''. The assignment of these labels was counterbalanced over the participants. The program recorded the response and its latency from the onset of the conclusion.

Results and discussion

Table 1 presents the percentages of correct responses and the mean latencies of all responses for the six sorts of valid inference. One participant failed to complete the experiment and his data have been excluded from the results.

Every participant was more accurate in evaluating the immediate conclusions than the iterative conclusions (Binomial test, $p = .5^{19}$). This difference was also reliable for each of the three sorts of problems based on different quantified premises (Binomial tests, $p = .5^{19}$, $p = .5^{18}$, $p = .5^{19}$, respectively). Similarly, 17 participants had faster latencies for the immediate conclusions than for the iterative ones, and only 2 participants had faster latencies for the iterative conclusions than for the immediate ones (Binomial test, $p = .0004$). The difference was also significant for each of the three sorts of problems (Wilcoxon test, $p < .01$, $p < .001$, $p < .001$, respectively).

TABLE 1
The percentages and the mean latencies (in seconds) for the experimental problems (with valid conclusions) in Experiment 1

Premises	*Immediate*		*Iterative*	
	Percent correct	*Mean latencies*	*Percent correct*	*Mean latencies*
1.	98	2.4	6	3.8
2.	92	3.2	8	4.8
3.	97	2.2	11	3.4
Overall	95	2.6	8	4.0

Percentages refer to correct responses only, whereas latencies refer to all the responses (correct and incorrect) pooled together.

As predicted, the immediate conclusions were rapidly and accurately evaluated, but the iterative conclusions took longer to evaluate and were usually rejected even though they were valid. The participants evidently had difficulty in making an iterative use of the quantified premise, and so they were unable to grasp the relation asserted in the iterative conclusions.

EXPERIMENT 2

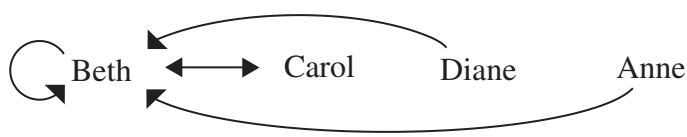
The inferences in the previous experiment were affirmative, that is, the specific premise asserted an affirmative relation that satisfied the antecedent possibility of the quantified premise. A contrasting negative inference can be made from a specific premise that asserts a negative relation:

Imagine a world in which there are four people: Anne, Beth, Carol, and Diane.
In that world, everybody loves anyone who loves someone.
Anne does not love Beth.
Does it follow that Beth does not love anybody? (Immediate conclusion)
Does it follow that Diane does not love Carol? (Iterative conclusion)

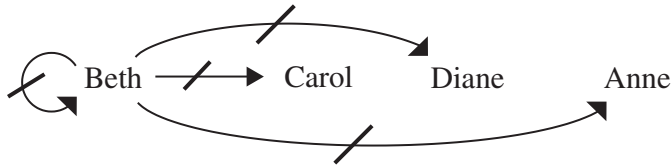
Previous studies of sentential and quantified reasoning have shown that reasoners spontaneously adopt different strategies of reasoning, although they are influenced by properties of the inferences (see, e.g., Bucciarelli & Johnson-Laird, 1999; Van der Henst, Yang, & Johnson-Laird, 2002). Reasoners are also likely to adopt various strategies in coping with iterative problems. One plausible strategy is to begin with a supposition:

Suppose that Beth loves someone, say, Carol

This supposition fits the antecedent possibility of the quantified premise (see 1 above). Hence, the consequent possibility follows, and everyone loves Beth:



But this possibility contradicts the specific premise that Anne does not love Beth, and so it cannot be the case that Beth loves anyone (the immediate conclusion):



The broken arrows represent the relation of *not* loving. Consider a further supposition:

Suppose that Diane loves someone, say, Anne.

It follows, as before, that everyone loves Diane. But, as we have already shown, Beth does not love Diane, because she does not love anyone. And so it cannot be the case that Diane loves anyone, and, in particular, Diane does not love Carol (the iterative conclusion). In sum, the quantified premise has the interesting consequence that in any finite world there are only two possibilities: either everyone loves everyone, or else noone loves anyone.

The model theory predicts that the two negative inferences should be harder than the two affirmative inferences. The negative inferences depend on grasping that a model is inconsistent with a premise, whereas the affirmative inferences follow merely from an iterative updating of models. Readers familiar with logic will notice the analogies between affirmative inferences and modus ponens, and between negative inferences and modus tollens. Modus ponens is easier than modus tollens (see Evans, Newstead, & Byrne, 1993, for a review of studies). The model theory accordingly makes two predictions: first, immediate conclusions should be easier than iterative conclusions; and, second, affirmative inferences should be easier than negative inferences. Experiment 2 tested these two predictions.

METHOD

Design. The participants acted as their own controls and carried out four main sorts of experimental inference based on whether the specific premise was affirmative or negative, and on whether the conclusion to be evaluated was an immediate or an iterative one. These four inferences depended on separate trials with two sorts of quantified premise:

(1) Everyone *verbs* anyone who *verbs* someone

and:

(2) No-one *verbs* anyone who does not *verb* someone.

TABLE 2
The eight sorts of valid inference in Experiment 2

Quantified premise	*Specific premise*	*Immediate conclusion*	*Iterative conclusion*
Everybody verbs anyone who verbs someone.	Anne verbs Beth.	Everyone verbs Anne.	Carol verbs Diane.
	Anne does not verb Beth.	Beth does not verb anybody.	Carol does not verb Diane.
Nobody verbs anyone who does not verb someone.	Anne does not verb Beth.	Nobody verbs Anne.	Carol does not verb Diane.
	Anne verbs Beth.	Beth verbs everybody.	Carol verbs Diane.

Table 2 summarises the resulting eight sorts of inference. There were eight filler trials in which the premises were combined with invalid conclusions, which were the negations of the corresponding valid conclusions. As in Experiment 1, the responses to invalid inferences were not comparable with responses to valid inferences. The inferences were presented in a different random order to each participant.

Participants. A total of 40 students (mean age 20.3 years) at the University of Milan volunteered to carry out the experiment, which lasted for about a quarter of an hour. None had taken any course in logic or in the psychology of reasoning.

Procedure and material. Each participant was given a 9-page booklet. The first page stated the instructions, and their key part read as follows:

> Imagine a world in which there are four persons: Anne, Beth, Carol and Diane. Each problem will assert two premises that are true in that world. ... After each pair of premises you will be given a conclusion. If the conclusion follows from the premises, check the "yes" box. Otherwise, check the "no" box. ... Give the responses that you think are correct, and think *as accurately as possible* ... and do not choose any answer by guessing.

In the actual problems, "verb" was replaced by one of the following: *appreciate, admire, esteem, respect*. The names and verbs in the premises and conclusions were randomised across the problems and across the participants, and no participant encountered more than one problem with a given set of names.

Results and discussion

Table 3 presents the percentages of correct conclusions (based on a total of 80 in each cell) for each of the four sorts of valid inference. There were no reliable differences between the two sorts of quantified premises, and so we

TABLE 3
The percentages of correct responses to the experimental problems in Experiment 2

Problems	*Immediate conclusion*	*Iterative conclusion*
Affirmative problems	79	24
Negative problems	36	6

have pooled the data from them. The results replicated those of the previous experiment and extended them to negative problems. Overall, 33 participants accepted more immediate conclusions than iterative conclusions, one participant accepted more iterative conclusions than immediate conclusions, and there were six ties (Binomial test, $p = .5^{23}$). Likewise, 27 participants accepted more affirmative conclusions than negative conclusions, 3 participants accepted more negative conclusions than affirmative conclusions, and there were 10 ties (Binomial test, $p = .5^{18}$). The difference in accuracy between the immediate and iterative conclusion was larger for the affirmative inferences than for the negative inferences (Wilcoxon test, $p < .05$). This interaction probably occurred because of a "floor" effect with negative iterative conclusions.

The results showed that the difficulty of iterative reasoning occurs with a variety of different logical structures. The phenomenon therefore probably depends on the need to use the quantified premise iteratively. The affirmative inferences in the present experiment appeared to be easier than those in Experiment 1. The effect, however, may have been a consequence of the difference in the instructions between the two experiments. In Experiment 1, the participants were told to respond as fast as possible, whereas in the present experiment they were told to respond as accurately as possible. Hence, the participants are likely to have re-examined the quantified premise after their initial conclusion, and so they had a better chance of noticing that it allowed an iterative inference. Nevertheless, the difficulty of negative iterative conclusions suggests that they are beyond the competence of logically naïve individuals.

EXPERIMENT 3

The previous experiments showed that inferences that call for an iterative use of a premise are more difficult than inferences that do not. According to the model theory (see the Introduction), the key steps for drawing an iterative conclusion are:

1 Everyone loves Anne. (The immediate conclusion from the initial model of the two premises)
2 Diane loves Anne. (From the same model)
3 Everybody loves Diane. (From an iterative use of the quantified premise)
4 Carol loves Diane. (From the same model).

The principal locus of difficulty in this sequence of steps is to update the initial model by making an iterative use of the quantified premise, i.e., the step that yield the conclusion that everyone loves Diane. Hence, if reasoning is based on extensional representations such as mental models, then a representation of "Everyone loves Anne" coupled with the knowledge that "Everyone" refers to Anne, Beth, Carol and Diane, yields the conclusion that Diane loves Anne with a minimum of effort. That is, this step is almost effortless in comparison with the iterative use of the quantified premise.

From a logical standpoint, the model theory predicts a chain of steps different from those of a derivation based on formal rules, such as those proposed by Rips (1994) and Braine and O'Brien (1998). We can summarise the steps in such a formal proof, as follows:

Since everyone loves Anne, everyone loves *someone*; given that everyone loves anyone who loves someone and everyone loves someone, it follows that everyone loves everyone, and so Carol loves Diane. The key steps in this formal proof are shown in the following *quantified* chain:

1 Everyone loves Anne. (The immediate conclusion).
2 Everyone loves someone.
3 Everyone loves everyone.
4 Carol loves Diane. (The iterative conclusion).

As far as we can tell, no existing psychological theory based on formal rules makes any strong prediction about whether naïve individuals should find the model chain more plausible, or less plausible, than the preceding quantified chain. Braine and O'Brien's formal rules do not deal with such inferences; Rips' system should deal with them, but the complete program implementing his system no longer exists (L. J. Rips, personal communication), and the surviving part is not able to prove the iterative inference. In contrast, the model theory predicts that naïve individuals should prefer the model chain to the quantified chain. It also predicts that the principal source of difficulty should be the iterative step illustrated above.

The present experiment tested these two predictions. The participants evaluated each of the four steps in the two contrasting chains of inference,

one based on the model theory, and the other using quantifiers according to formal logic. The goal was to test which chain was easier to validate, which chain yielded more correct evaluations of the final iterative conclusion, and which steps in the preferred chain were hardest. The first and fourth steps in the two chains were, of course, identical.

Method

Design and materials. The participants acted as their own controls and carried out two versions of each of three problems. Each chain began with the immediate conclusion and ended with the iterative conclusion, but in between there were two intermediate conclusions. In the *model* chain, the intermediate conclusions referred to specific individuals, for example:

Diane loves Anne.
Everyone loves Diane.

In the *quantified* chain, based on formal rules of inference, there were intermediate conclusions containing quantifiers (corresponding to temporary names), for example:

Everyone loves someone.
Everyone loves everyone.

The complete set of problems and their corresponding chains of inference are summarised in Table 4. Each problem had a different content based on the verbs: *appreciate, admire, esteem, respect, love, trust*, and the problems were presented in a different random order to each participant.

Participants. A total of 40 students (mean age 22.1 years) of the University of Trento volunteered to take part in the experiment, which lasted for about 15 minutes. None had taken any course in logic or the psychology of reasoning.

Procedure. Each participant was given a seven-page booklet and the first page stated the instructions. The key instructions translated from the Italian were as follows:

> After each pair of premises you will be given four conclusions. If a conclusion follows from the premises, check the "yes" box by it. Otherwise, check the "no" box. ... Please solve the problems in the order in which they have been given to you. ... Give the responses that you judge to be correct thinking *as accurately as possible*, ... and do not respond randomly.

TABLE 4
Examples of the two sorts of problem in Experiment 3

Premises	*A model chain*	*A quantified chain*
Everyone loves anyone who loves someone. Anne loves Beth.	Everyone loves Anne. Diane loves Anne. Everyone loves Diane. Carol loves Diane.	Everyone loves Anne. Everyone loves someone. Everyone loves everyone. Carol loves Diane.
No-one loves anyone who does not love someone. Anne does not love Beth.	No-one loves Anne. Diane does not love Anne. No-one loves Diane. Carol does not love Diane.	No-one loves Anne. Everyone does not love someone. No-one loves anyone. Carol does not love Diane.
Anyone who is loved loves everyone. Anne loves Beth.	Beth loves everyone. Beth loves Carol. Carol loves everyone. Carol loves Diane.	Beth loves everyone Everyone is loved. Everyone loves everyone. Carol loves Diane

TABLE 5
The mean numbers of correct responses (acceptances of conclusions as valid out of three) for each conclusion in the two sorts of chains of inference in Experiment 3

Conclusions	*Model chain*	*Quantified chain*
1. Immediate conclusion	2.3	2.1
2. Intermediate 1	2.3	1.1
3. Intermediate 2	0.9	0.3
4. Iterative conclusion	1.0	0.5

Results and discussion

Table 5 presents the mean number of correct acceptances (out of a maximum of three) for each conclusion in the two sorts of chains of inference. We pooled the results for the three pairs of problems, because there were no reliable differences among them. There was no significant difference in the number of participants giving the correct answer to the identical *immediate* conclusion at the start of the two sorts of chain (4 participants accepted the conclusion in the model chain more often than the quantified chain, and the remaining 36 participants were ties). But the participants accepted the remaining conclusions more often in the model chain than in the quantified chain: for the first intermediate conclusion, 28 participants accepted the model conclusion more often than the quantified conclusion, and the rest were ties (Binomial test, $p = .5^{28}$); for the second intermediate conclusion, 16 participants accepted the model conclusion more often than the quantified conclusion and the rest were ties (Binomial test, $p = .5^{16}$); and for the identical iterative conclusion, 16 participants accepted the conclusion in the model chains more often than the conclusion in the quantified chains, 4 participants accepted the quantified conclusion more often than the model conclusion, and the rest were ties (Binomial test, $p = .006$).

In the model chain, there was no significant difference in the number of participants accepting the immediate conclusions and the first intermediate conclusions (all 40 participants were ties). But 32 participants accepted the first intermediate conclusions more often than the second intermediate conclusions, and the rest were ties (Binomial test; $p = .5^{32}$). There was no significant difference between the number of participants accepting the second intermediate conclusion and the iterative conclusion.

In the quantified chain, 24 participants accepted the immediate conclusions more often than the first intermediate conclusions; the rest were ties (Binomial test, $p = .5^{24}$). Sixteen participants accepted the first intermediate conclusions more often than the second intermediate conclusions, four accepted the second intermediate conclusions more often than the first, and the rest were ties (Binomial test, $p = .006$). Eight participants

accepted the iterative conclusion more often than the second intermediate conclusion, and 32 were ties (Binomial test, $p = .5^8$).

In sum, an iterative conclusion was more acceptable following a chain based on the model theory than following a different but equally valid chain, based on formal rules of inference. Likewise, the intermediate conclusions were more acceptable in the model chain rather than in the quantified chain. The results also corroborated the prediction that the main locus of difficulty is inferring the second intermediate conclusion. This step is the only one that requires updating the initial model. In contrast, difficulties arose in all the intermediate steps in the quantified chain—a phenomenon that suggests that this chain was unnatural for naïve reasoners. A significant number of participants accepted the iterative conclusion in the quantified chain, but not the second intermediate conclusion. This result implies that these participants may have reached the iterative conclusion by a different path.

The experiment corroborates the model theory's predictions. Naïve reasoners prefer a sequence of inferential steps that concern individuals to a sequence that contains quantifiers of the sort that occur in logical proofs. Most participants behaved *as if they were* building an initial model of the premises, and then updated it iteratively. Only a minority of participants considered the two derivations equally difficult. Of course, the results do not eliminate the possibility that reasoners rely on formal rules, but current theories based on formal rules do not appear to predict them.

EXPERIMENT 4

The previous experiments showed that naïve reasoners do not readily make iterative inferences, and that the difficulty is predicted by the model theory. But what sort of strategies do naïve reasoners tend to adopt when they tackle iterative inferences? In order to try to answer this question, the participants in Experiment 4 had to write down their explanations for why the conclusions of iterative inferences followed from the premises. The model theory predicts that their protocols should reveal the use of mental models representing individuals rather than quantifiers.

Method

Design. The participants carried out five problems, each presented with a single conclusion that they had to justify. Three affirmative problems were from Experiment 1 and each was coupled with an iterative conclusion; and two negative problems were from Experiment 2, one coupled with an immediate conclusion and the other coupled with an iterative conclusion. The problems are presented in Table 6. Their contents and order were the

TABLE 6
The five problems in Experiment 4 with their particular contents and in their order of presentation

Premises	*Conclusion*
(1) Everybody esteems anyone who esteems someone. Anne esteems Beth.	Diane esteems Carol.
(2) Carol respects all those who respect Diane. Diane respects herself.	Carol respects herself.
(3) Anyone who is appreciated appreciates everybody. Anne appreciates Beth.	Carol appreciates Diane.
(4) Everybody admires anyone who admires someone. Anne does not admire Beth.	Beth does not admire anyone.
(5) Everybody loves anyone who loves someone. Anne does not love Beth.	Carol does not love Diane.

same for each participant in order to keep the task as similar as possible across the participants. The negative problems were presented last in order to try to improve performance with them.

Participants. A total of 14 students (mean age 22.2 years) at the University of Milan volunteered to participate in the experiment, which took about 1 hour. None had taken any course in logic or in the psychology of reasoning.

Procedure and material. Each participant was given the instructions and the problems in a booklet. The key instructions translated from the Italian were as follows:

> Your task is to try to understand why the conclusion follows from the premises, writing down your arguments and thoughts. We are interested in how people tackle these problems, and so the more you write the better. If you wish you can use drawings, diagrams, graphs, or any other written means that can help to clarify your thinking. You can go back and forth between the problems, but please note down when you do so.

Results and discussion

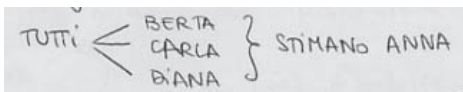
The protocols showed that the majority of the participants explained the validity of the iterative conclusions of the affirmative problems in the way predicted by the model theory. Here are two typical protocols (both for the first problem):

If from premise 1 we see that everybody esteems anyone who esteems someone, and knowing that Anne esteems Beth, it follows that everyone esteems Anne, or, more properly, that Beth, Carol, and Diane esteem Anne. Therefore if Carol esteems Anne, everybody esteems Carol. From this it follows that Diane esteems Carol.

The first impression is that the conclusion does not necessarily follow from the premises. What immediately follows is that Anne is esteemed by everybody, inasmuch she esteems Beth. In the first premise it is definitely not said that everybody should esteem anyone, but more simply that those who esteem someone are themselves esteemed. Anne esteems, and therefore she is esteemed by Beth, Carol and Diane. If Carol esteems Anne she is herself esteemed by everybody, and thus also by Diane. Therefore Diane esteems Carol.

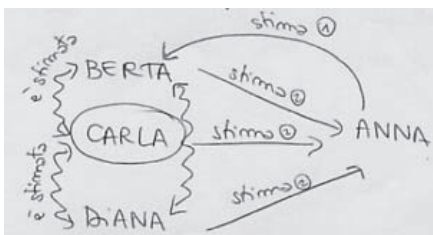
The protocols contained five principal features that are relevant to the model theory, and we will describe each of them in turn. As the preceding protocols illustrate, the participants tended to start by drawing the immediate conclusion corresponding to the initial model of the two

The conclusion does not follow necessarily, because if I assume premise 1, I know for sure that a person is esteemed if she esteems someone. Hence, if I knew that Carol esteemed someone, I could definitely accept the conclusion, but given that in the premises nothing explicit is told about the fact that Carol esteems or does not esteem someone, the conclusion could follow but does not follow necessarily. From premise 2 I reckon that:



(everybody) (Beth, Carol, Diane) (esteem Anne)

But I do not find any direct relation between Diane and Carol.
I understood the solution after solving the third problem. It is:



But then Carol esteems Anne, and hence she is esteemed by everybody and also by Diane; so Diane esteems Carol

("stima" = esteems; "è stimata" = is esteemed)

1 is the premise 2.
2 follows directly from premise 2.
The waved arrows follow indirectly from premise 2.
The italicised translations under the diagrams were not in the original.

Figure 1. A typical protocol with original diagrams (for the first problem in Experiment 4).

premises: they did so in 37 out of the 42 protocols from affirmative problems, and all 14 participants did so on more than half of the trials (Binomial test $p = .5^{14}$). Second, the participants used the quantified premise iteratively, i.e., they specifically referred to this operation and often they drew a diagram: 34 out of the 42 affirmative protocols, and 13 out of the 14 participants made an explicit iterative step on more than half the trials (Binomial test, $p = .0009$). Third, in 16 out of the 42 protocols, the participants included diagrams of the iterative step. Figure 1 presents an example of such a protocol. Fourth, the participants gave a correct explanation of the iterative conclusion on 32 out of the 42 trials, and 13 out of the 14 participants generated correct explanations on more than half the trials (Binomial test $p = .009$). Fifth, the protocols tended to refer to specific individuals in intermediate conclusions. Problem 2 has a quantified premise that mentions two specific individuals, and so we excluded it from this analysis. But 22 out of the remaining 28 protocols contained intermediate conclusions referring to specific individuals; and 11 out of the 14 participants used intermediate conclusions referring to specific individuals more often than intermediate conclusions referring to quantifiers, while three participants did the opposite (Binomial test, $p = .03$).

The protocols from negative problems also corroborated the model theory, and illustrated many of the same features. Problem 4 had an immediate conclusion, as shown here:

> Everybody admires anyone who admires someone
> Anne does not admire Beth
> Does it follow that Beth does not admire anyone?

With this problem, the participants tended to start by integrating the two premises. Here is a typical example:

> ~~From premises 1 and 2 follows that everybody does not admire Anne and that therefore Beth, Carol and Diane *do not* admire Anne. From premises 1 and 2 we see that~~ If Beth admired someone, she would be admired by everyone, including Anne. Given that Anne does not admire Beth, it follows that Beth does not admire anyone.

This participant crossed out the first part of her protocol, which shows that she began with an inference based on a misreading of the quantified premise. A few other participants also misinterpreted the quantified premise. But, as in the example, half of the participants noticed the contradiction between Beth admiring someone and Anne not admiring her, and accordingly gave a correct explanation of the conclusion (see the introduction to Experiment 2 for a model-based account of how this contradiction is discovered).

Problem 5 was a negative problem with an iterative conclusion:

> Everybody loves anyone who loves someone
> Anne does not love Beth
> Does it follow that Carol does not love Diane?

Here is a protocol illustrating a sequence of reasoning leading to a correct explanation of the inference:

> Beth is not loved by Anne; this means that she does not love anyone. If only she loved someone, she would be loved herself. Thus Beth does not love Anne, Carol, and Diane. However, if she does not love them, it is because they themselves do not love anyone. If only they loved someone, they would be loved. Therefore, Carol, who does not love anyone, does not love Diane.

As the protocol illustrates, many participants began by drawing the immediate conclusion (9 out of the 14 participants). Most of them then drew a sequence of intermediate conclusions about specific individuals, although some also drew intermediate quantified conclusions. In keeping with the difficulty of the problem, only a minority of participants explicitly mentioned an iterative use of the quantified premise (five participants) and gave a correct explanation of the conclusion (six participants).

Not surprisingly, the participants performed much better in the present experiment than in Experiment 2. In the present experiment, they *knew* that the conclusions were valid. Hence, if they used an argument that failed to yield the conclusion, they knew that they had gone wrong, and they searched for a different argument. In addition, the inferences were presented in a helpful order that started with the easiest inferences and ended with the hardest inferences.

As in other sorts of reasoning, the participants used several different strategies for explaining the five inferences, and the details of their individual protocols showed considerable variation in particular inferential steps (see, e.g., Van der Henst et al., 2002). Nevertheless, there were some basic steps in common. The participants tended to draw the appropriate immediate conclusion before they drew an iterative conclusion. Likewise, as the model theory predicted, their intermediate conclusions tended to concern specific individuals rather than to use quantified expressions. Only a few participants could correctly explain the iterative conclusion to the negative inference.

GENERAL DISCUSSION

Reasoning is poorer when inferences call for an iterative use of a quantified premise (Experiment 1), and negative iterative conclusions are overwhelmingly difficult for logically naïve individuals (Experiment 2). When such reasoners have to evaluate a chain of inferences leading from an immediate to an iterative conclusion, they perform more accurately with a chain of model-based inferences than with a chain of quantified inferences. The only

difficult step in the model-based chain is the one requiring the iterative update of the initial model (Experiment 3). When reasoners have to explain the validity of iterative inferences, they can do so in the case of affirmative inferences, but they have some difficulty with negative inferences (Experiment 4). They again show a bias towards intermediate conclusions concerning specific individuals.

Reasoners could in principle consider the iterative consequences of the quantified premise alone. Hence, given a pair of premises, including the quantified premise:

Everyone loves anyone who loves someone.

they could infer:

If someone loves someone then everyone loves someone.

from which it follows:

If someone loves someone then everyone loves everyone.

They could then combine this conclusion with the specific premise:

Anne loves Beth

in order to infer that everyone loves everyone. Both the immediate conclusion (Everyone loves Anne) and the iterative conclusion (Carol loves Diane) follow at once. In other words, a logically feasible strategy is to update the model of the quantified premise iteratively *before* it is combined with the specific premise. The same strategy could in principle be used with negative inferences. No participant in Experiment 4 ever used this strategy. Indeed, the protocols suggest that many participants failed to infer that a consequence of one person loving another is indeed that everyone loves everyone. Evidently, logically naïve individuals do not normally make hypothetical iterations of quantified premises.

The iterative use of a premise is difficult, and so it is hard to draw iterative conclusions. The model theory predicts the phenomenon because reasoners need to update their model of the premises. Theories of reasoning based on formal rules of inference could be formulated to make the same prediction, but current formal rule theories have some difficulty in accounting for the phenomena. Mental models, however, represent individuals, not variables, and chains of inference in which intermediate conclusions refer to individuals are evaluated more accurately than chains in which intermediate conclusions refer to quantified variables.

Although the quantified premises used in our experiments are complicated, artificial, and rarely encountered in daily life, they are useful in the study of human reasoning, because they clarify one of its central features. Naïve reasoners prefer to think about individuals rather than variables. In our experiments, they did not expand the model of the quantified premise by itself, but integrated it with the specific premise. Super-intelligent entities would grasp at once the iterative consequences of a quantified assertion. They would immediately see that as soon as one person loves another, everyone loves everyone. The anthropologist Claude Levi-Strauss once claimed that a barbarian is a person who believes in barbarism (see Gellner, 1970). That is, any person who believes that other people are barbarians is a barbarian too. There are persons with such a belief, i.e., they believe that other people are barbarians. Hence, they are barbarians. And we believe that there are such individuals. It follows that we too are barbarians. As Gellner comments, this definition of barbarian "spreads barbarism like wildfire through the mere awareness of it" (p. 31). Yet the iterative consequence of the definition is not obvious. Unlike super-intelligent entities, human reasoners do not tend to make a fully iterative use of quantified assertions. Their preferred strategy is to integrate a quantified premise with a specific premise, and then perhaps to update the integrated model iteratively. This bias is in keeping with the general principle of the model theory: models represent specific individuals rather than variables.

Manuscript received 23 December 2002
Revised manuscript received 24 July 2003

REFERENCES

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

Bub, D., & Gum, T. (1991). *Psychlab experimental software*. Montreal: McGill University.

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

Cherubini, P., Garnham, A., Oakhill, J., & Morley, E. (1998). Can any ostrich fly? Some new data on belief bias in syllogistic reasoning. *Cognition*, *69*, 179–218.

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

Gellner, E. (1970). Concepts and society. In B. R. Wilson (Ed.), *Rationality* (pp. 18–49). Oxford: Blackwell.

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

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

Oakhill, J., Garnham, A., & Johnson-Laird, P. N. (1990). Belief bias in syllogistic reasoning. In K. Gilhooly, M. Keane, R. Logie, & G. Erdos (Eds.), *Lines of thinking* (Vol. I). New York: Wiley.

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

Van der Henst, J-B., Yang, Y., & Johnson-Laird, P. N. (2002). Strategies in sentential reasoning. *Cognitive Science*, *26*, 425–468.