Memory & Cognition
2003, 31 (7), 1105-1113

Counterexamples in sentential reasoning

P. N. JOHNSON-LAIRD and URI HASSON
Princeton University, Princeton, New Jersey

How do logically naive individuals determine that an inference is invalid? In logic, there are two ways to proceed: (1) make an exhaustive search but fail to find a proof of the conclusion and (2) use the interpretation of the relevant sentences to construct a counterexample—that is, a possibility consistent with the premises but inconsistent with the conclusion. We report three experiments in which the strategies that individuals use to refute invalid inferences based on sentential connectives were examined. In Experiment 1, the participants' task was to justify their evaluations, and it showed that they used counterexamples more often than any other strategy. Experiment 2 showed that they were more likely to use counterexamples to refute invalid conclusions consistent with the premises than to refute invalid conclusions inconsistent with the premises. In Experiment 3, no reliable difference was detected in the results between participants who wrote justifications and participants who did not.

Logically speaking, an inference is either valid or invalid. A conclusion of a valid inference—a *valid* conclusion, for short—is one that is necessarily true given that the premises of the inference are true. A conclusion of an invalid inference, an *invalid* conclusion, is not necessarily true given that the premises are true (see, e.g., Jeffrey, 1981). Individuals with no training in logic are able to draw valid conclusions (see, e.g., Evans, Newstead, & Byrne, 1993). They are also able to reject invalid conclusions. How do they do it? In logic, one method is to rely on formal rules of inference and to make an exhaustive search for a proof in which the conclusion is derived from the premises. If the search fails, the inference is invalid. Another method is to consider the interpretation of the premises and to search for a counterexample—that is, a possibility consistent with the premises but inconsistent with the conclusion. However, as logicians such as Quine (1974) and Barwise (1993) have pointed out, if you use the first method and search for a formal proof, you cannot know for certain that your conclusion is invalid; you may have overlooked a proof. In contrast, a counterexample is a manifest demonstration of invalidity, and both formal and semantic methods exist in logic for searching for counterexamples (see Jeffrey, 1981).

A counterexample to a generalization is an instance to the contrary—for example, a black swan falsifies the claim that all swans are white (see, e.g., Holyoak & Glass, 1975). Likewise, a counterexample to an inference is a possibility to the contrary. It establishes that a conclusion fails to follow validly from the premises, because it is a possibility in which the premises are true but the conclusion is false. But do counterexamples play any role in the reasoning of logically naive individuals? Some psychologists have argued that they do not. For example, they are not part of Rips's (1994) PSYCOP account of reasoning, which is based on formal rules of inference. Similarly, Polk and Newell (1995) have argued that their model-based theory of reasoning provides a good fit with their data on individual differences without there being a need for a parameter value reflecting the use of counterexamples. Reasoning, they claim, is verbal comprehension, and "not the construction of alternative models to falsify a putative conclusion" (p. 557). Yet the case against counterexamples is not decisive, and the present article in turn confronts it with systematic counterexamples.

According to the mental model theory of reasoning (e.g., Johnson-Laird & Byrne, 1991), individuals establish that a conclusion follows validly from premises by determining that it holds in all the possibilities consistent with the premises. Reasoners represent each of these possibilities in a mental model. The more models that they have to construct in order to draw a valid conclusion, the harder the inference is—that is, it tends to take longer and be more error prone (Johnson-Laird & Byrne, 1991). When individuals find a model of a situation in which the premises are true but the conclusion is false, they judge that the conclusion is invalid. In this case, they may still infer that the conclusion is probable (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999) or at least possible (Bell & Johnson-Laird, 1998), but they know that it is not necessarily true, because they have discovered a counterexample to it.

An invalid conclusion, as we noted earlier, is one that is *not* necessarily true given that the premises are true. It follows that conclusions can be invalid in two different ways. On the one hand, a conclusion can be invalid because it is inconsistent with the premises—that is, it does

This research was supported by Grant BCS-0076287 from the National Science Foundation to study strategies in reasoning. We thank our colleagues Ruth Byrne, Sam Glucksberg, Geoffrey Goodwin, Catrinel Haught, and Susanna Reynolds for advice. We also thank Jonathan Evans, Mike Oaksford, and an anonymous reviewer for their helpful suggestions. Correspondence concerning this article should be addressed to P. N. Johnson-Laird, Department of Psychology, Princeton University, Princeton, NJ 08544 (e-mail: phil@princeton.edu).

not hold for any of the possibilities compatible with the premises. The following inference illustrates this sort of invalidity in which the conclusion is inconsistent with the premise:

(1) Either Dan is in Madrid or else Bill is in Seoul, but not both.

Therefore, Dan is in Madrid and Bill is in Seoul.

The premise is consistent with two possibilities (shown here in full on separate lines):

Dan–Madrid	¬Bill–Seoul
¬Dan–Madrid	Bill–Seoul

where "Dan–Madrid" denotes Dan's being in Madrid, "¬" denotes negation, and so "¬ Dan–Madrid" denotes Dan's not being in Madrid. Each of these possibilities is a counterexample to the conclusion: It is impossible that Dan and Bill are at the locations in the conclusion. The conclusion is inconsistent with the premise, and so it cannot follow from it. On the other hand, a conclusion can be invalid but, nevertheless, consistent with the premises. In this case, it holds for at least one possibility compatible with the premises, but not for all such possibilities. The following inference illustrates this sort of invalidity:

(2) Dan is in Madrid or Bill is in Seoul, or both.

Therefore, Dan is in Madrid and Bill is in Seoul.

The premise is consistent with three possibilities (shown here in full on separate lines):

Dan–Madrid	¬Bill–Seoul
¬Dan–Madrid	Bill–Seoul
Dan–Madrid	Bill–Seoul

The conclusion is consistent with the premise: It corresponds to the third of these models. Hence, given the truth of the premise, the conclusion could be true but it need not be. The first and second models of the premises are both counterexamples in which it fails to hold. Reasoners who can build a model of one of these possibilities and recognize it as a counterexample know that the inference is invalid: The premises can be true, and the conclusion can be false. And so the conclusion is not necessarily true. The number of models compatible with the premises should not directly affect the difficulty of establishing invalidity. Reasoners do not need to take all of them into account, and indeed, they may discover a counterexample in the first model that they construct, especially in the case of inconsistent conclusions for which any model serves as a counterexample. However, the theory does yield a prediction distinguishing between the two sorts of invalid conclusion. The invalidity of a conclusion inconsistent with the premises should be easier to establish than the invalidity of a conclusion consistent with the premises. Any model of the premises refutes a conclusion inconsistent with them, but not any model of the premises refutes a conclusion consistent with them. Suppose, as often happens, reasoners fail to consider all the models of possibilities compatible with the premises (see, e.g., Evans, Handley, Harper, & Johnson-Laird, 1999; Johnson-Laird & Byrne, 1991). They are still likely to evaluate an inconsistent conclusion correctly, because it conflicts with all the models of the premises. But they are at risk of erring with a consistent but invalid conclusion, because they may overlook a model that is a counterexample to the conclusion. The model theory therefore predicts that inconsistent conclusions should yield a greater number of correct evaluations than would consistent but invalid conclusions. Braine and O'Brien (1991) defended the use of formal proofs to establish invalidity if the conclusion is inconsistent with the premises. But if it is consistent with the premises, Braine and O'Brien suggested that mental models might be used to establish its invalidity. Counterexamples, however, can be used to refute either sort of invalid conclusion. So, do logically naive individuals use counterexamples at all?

Most psychological studies of reasoning have called for participants to judge the validity of given conclusions, to choose the valid conclusions from multiple choices, or to draw their own conclusions. Unfortunately, none of these methods can reveal how reasoners establish invalidity. For that, experimenters must use a different sort of procedure. Hence, Bucciarelli and Johnson-Laird (1999) video-recorded their participants as they constructed external models of syllogisms, such as

(3) Some of the chefs are musicians.

All the painters are musicians.

Therefore: Some of the chefs are painters.

The participants were able to construct external models to refute invalid conclusions, such as the one in the preceding example. They also appeared to consider counterexamples when they inferred their own spontaneous conclusions. In contrast, Newstead, Handley, and Buck (1999) reported a study in which participants drew diagrams corresponding to putative syllogistic conclusions. Contrary to expectation, those participants who constructed more distinct alternative diagrams were no more accurate in their reasoning than those who constructed fewer distinct alternative diagrams. Even when the participants were told that such diagrams could be used to solve syllogisms (in Experiment 4), there was no reliable correlation between the number of diagrams they drew and the accuracy of their responses. In a recent study, however, Newstead, Thompson, and Handley (2002) found that a measure of individual differences, which reflected the ability to generate alternative possibilities, did correlate with participants' ability to refute invalid syllogistic conclusions. In fact, with syllogisms, reasoners tend to construct only a single model, and so their performance is always better with those syllogisms that have only a single model than with those that have multiple models (see, e.g., Bucciarelli & Johnson-Laird, 1999; Evans et al., 1999). Indeed, Newstead et al. (1999) report that fewer than 20% of the responses in their Experiment 3 were correct. Hence, with hindsight, syllogisms are not the best inferences with which to investigate counterexamples. Naive reasoners tend to misunderstand syllogistic premises, overlooking alternative possibilities consistent with them, and so the results yield systematic underestimates of the use of counterexamples (as Roberts, in press, has pointed out).

Studies of counterexamples have focused on syllogisms, and for the reasons that we outlined, their results have not been clear-cut. We accordingly decided to investigate the use of counterexamples in the domain of sentential reasoning—that is, reasoning that hinges on the use of negation and sentential connectives, such as "if," "or," and "and." We do not presuppose that these connectives have the same meanings as they do in sentential logic, but with simple neutral sentences about people in cities—for example, "Either Dan is in Madrid or else Bill is in Seoul, but not both"—there is a close correspondence (Barres & Johnson-Laird, 2003). By *neutral* here, we mean that neither the meaning of the assertion nor general knowledge modulate its interpretation (see Johnson-Laird & Byrne, 2002). Given neutral assertions of the form *A and B*, individuals list as compatible with the truth of the assertion only one possibility in which both *A* and *B* occur. Given neutral assertions of the form *Either A or else B, but not both*, individuals list two possibilities as compatible with the truth of the assertion: one in which *A* alone occurs and one in which *B* alone occurs. Given assertions of the form *A or B, or both*, individuals list the two previous possibilities and a third one in which both *A* and *B* occur. Given assertions of the form *If A then B*, individuals list at least two possibilities as compatible with the truth of the assertion—*A* and *B*, *not-A* and *not-B*—and some individuals also list *not-A* and *B* (see Barrouillet & Leças, 1999). These possibilities correspond to the logical interpretation of the connectives. In sentences that are not neutral, however, natural language departs from logic (see, e.g., Johnson-Laird & Byrne, 2002).

Experimenters working with sentential connectives are faced with a dilemma: Inferences need to be simple enough for naive reasoners to respond correctly, but such inferences elicit responses that are too rapid for observers to determine the nature of the underlying processes (cf. Van der Henst, Yang, & Johnson-Laird, 2002). We accordingly gave our participants quite simple inferences, but in order to discover whether they grasped the force of counterexamples, we asked them to write down justifications for their responses. This procedure had a twofold advantage. First, it enabled us to check how the participants had understood the sentential connectives in the problems and, in this way, to overcome the endemic weakness in studies of syllogisms. Second, it enabled us to determine the strategies that the participants relied on to justify their evaluations of conclusions as valid or invalid. The model theory predicts that naive reasoners are likely to develop different strategies, and evidence has corroborated this claim for valid conclusions (see, e.g., Van der Henst et al., 2002). The theory also predicts that reasoners should rely on counterexamples, but we were curious to discover whether they would make use of other strategies of refutation. The demand for a justification may, of course, elicit strategies that would not be used in other circumstances, and so we must evaluate whether this task prompts different response patterns from those that would be found otherwise. Our principal goal, however, was to determine whether or not naive reasoners grasp the force of counterexamples.

EXPERIMENT 1

Method

Participants. In Experiment 1, we tested 25 Princeton undergraduates, who participated either for payment or to fulfil a course requirement. None of them had taken any courses on logic or cognate topics.

Design. The participants acted as their own controls and evaluated nine pairs of premises presented once with a valid conclusion and once with an invalid conclusion, but with different contents. The invalid conclusions were all consistent with the premises but did not follow from them necessarily. Table 1 presents the forms of

Table 1
The Forms of Problem in Experiment 1, the Percentages of Correct Evaluations of the Invalid Conclusions, and the Percentages of Counterexamples Used in the Correct Evaluations

Problem	Form	Percentage of Correct Evaluations	Percentage of Counterexamples Given a Correct Evaluation
1	A. B or C. ∴ C and A	70	83
2	A or B. B or else C. ∴ C if and only if A.	53	88
3	A or B. B or else C. ∴ C and A.	47	55
4	A if and only if B. C or B. ∴ A or else C.	50	87
5	A and B. B or C. ∴ If A then C.	47	33
6	A or else B. B or else C. ∴ B.	76	84
7	If A then not B. B or C. ∴ A or C.	68	100
8	If A then B. If B then C. ∴ C.	90	17
9	If A then not B. If C then B. ∴ A or else C.	23	75
	Overall	55	70

these problems. The 18 different problems were presented in two blocks separated by a brief interval; each block contained an approximately equal mixture of valid and invalid problems, and the order of the problems was randomized for each participant.

Materials and Procedure. The contents of the problems paired names of individuals with locations, as in Example 2 above, and each combination of names and locations was used once in the experimental session. The participants were tested individually. The key instructions were the following: "Your task is to determine whether the conclusion necessarily follows from the premises, that is, given that the premises are true, is the conclusion bound to be true too? . . . Imagine that you're explaining to someone why the conclusion does, or does not, follow from the premises. You must write down after each response, the reason that you have made it."

Results and Discussion

We scored a response as correct only when both its evaluation and its justification were correct, and we rejected the data from 3 of the 25 participants because they misunderstood the task as asking whether a conclusion was "possible." Table 1 presents the percentages of correct evaluations of each of the invalid inferences and the percentages of counterexamples. Table 2 describes the criteria that distinguish the three main classes of strategy and gives verbatim protocols for each sort of strategy (see Experiment 2 for an assessment of the reliability of the classification). On 70% of the trials in which the participants correctly rejected an invalid conclusion, they used a counterexample—that is, they justified their evaluation with an explicit description of a possibility in which the premises were true but the conclusion was false. On the remaining 30% of evaluations of invalidity, the participants used various miscellaneous strategies (see Table 2 for the breakdown of percentages within this category). Every participant used counterexamples on at least two of the nine invalid inferences (Binomial test, $p = .5^{22}$, assuming a prior probability of .5 for the use of the strategy). All but one of the problems elicited counterexamples as the preferred strategy (Binomial test, $p < .025$, again assuming a prior probability of .5 for the use of the strategy). Most important, the use of counterexamples correlated with correct evaluations. We assessed this relation by considering the proportion of correct evaluations on which each participant used a counterexample and the total number of correct evaluations that the participant made. Given that all the participants produced at least one correct evaluation, these two variables are, in principle, orthogonal—for example, a participant could have evaluated all 20 problems correctly but never have used a counterexample. In fact, there was a reliable correlation between the two variables (Pearson's $r = .44$, $p < .05$).

For problems with valid conclusions, the participants answered correctly 76% of the time. When these problems were wrongly misrecognized as invalid, the errors were made on the basis of the claim that there was no dependency between the two constituents mentioned in the conclusion. Not surprisingly, counterexamples were never used to refute such conclusions, since no counterexamples exist.

Table 2
The Three Main Classes of Strategy

The Counterexample Strategy

Criteria: Participants explicitly described a possibility in which the premises were true but the conclusion was false.

Example 1 (from Problem 3, Table 1): "No, it could be the case that Tom is in Paris [B], without Vern being in Tokyo [A] or Jill being in Lima [C]."

Example 2 (from Problem 7, Table 1): "No, Abe can be in Belgrade [B] and Kyle not in Dubai [not-A] and Seth not in Moscow [not-C] all at the same time."

The Contradiction Strategy

Criteria: Participants explicitly state that given the premises, the conclusion is impossible, contradicts the premises, or is inconsistent with them.

Example 1 (from Problem 1, Table 2): "No. If Yan is not in Paris [not-A] then Gail cannot be in Rome [not-C]."

Example 2 (from Problem 3, Table 2): "No, If we did know that Clint was not in Riga [not-A] then Lans must be in Bradford [C]."

Miscellaneous Strategies and Their Percentages of Occurrence in Experiments 1 and 2

Metalogical (7% in Experiment 1, 7% in Experiment 2)

Criteria: Participants appealed to logical principles unrelated to the specific content of the premises.

Example (from Problem 8, Table 1): "It could be that Nick is in New York [C], but nothing definite can follow from two conditional statements."

Completion to Truth (6% in Experiment 1, 6% in Experiment 2)

Criteria: Participants contrasted a self-generated valid conclusion with the given invalid conclusion.

Example (from Problem 7, Table 1): "No, the conclusion shouldn't be an "or" conclusion. If Kyle is in Dubai [A] THEN Seth is in Mecca [C] for it to be Yes."

Dependency (7% in Experiment 1, 22% in Experiment 2)

Criteria: Participants identified a missing premise required for the validity of the conclusion.

Example (from Problem 8, Table 1): "There needs to be a prior instruction for Platt to be in Cairo [C] because his being there relies on other factors (Lans [A] + Neils [B])."

Noted Possibility (9% in Experiment 1, 5% in Experiment 2)

Criteria: Stated that the conclusion is possible, but not necessary.

Example (from Problem 3, Table 1): "Tom can be in Paris [B], but we can't tell for sure."

Note—The table states the criteria used by the judges to classify the justifications in Experiments 1 and 2, and two examples of the two main strategies and one example of each of the miscellaneous strategies. The table shows the percentages of use of each strategy in the miscellaneous class for Experiments 1 and 2. The percentages of use of the counterexample strategy in Experiment 1 are presented in Table 1; the percentages of use of the counterexample and the contradiction strategies in Experiment 2 are presented in Table 4.

The results corroborated the use of counterexamples to justify judgments of invalidity. They also showed, however, that reasoners use other strategies to establish invalidity. An inference of the form *A and B*; *B or C*; *therefore If A then C* (Problem 5 in Table 1) tended to elicit the strategy in which reasoners noted that there was no dependency between the propositions in the conditional conclusion—for example, *A is true, but just because B doesn't mean C*. One reason that individuals may not have used an explicit counterexample in this case is that they may have envisaged the mental models of the premises:

A B
A B C

The first model is a counterexample, but individuals, rather than highlighting its role, which is implicit because it depends on the *lack* of C in the model, may instead have noted that that there is no dependency between A and C, because A can occur without C. With a more lenient scoring system, this response might have been scored as an implicit use of a counterexample. The problem that elicited the lowest proportion of counterexamples was one in which the two premises were conditional statements and the conclusion was a categorical statement (Problem 8 in Table 1). Some participants rejected the conclusion on the "higher" principle that nothing definite follows from two conditionals (an incorrect principle—e.g., if A then B, if not-A then B implies B). This claim could have been supported by the mental models of the premises:

A B C
. . .

Other problems (e.g., Problem 6 in Table 1) also have definite conclusions but elicited a large percentage of counterexamples. This pattern suggests that the form of the premises affects the use of counterexamples. Conditional premises, as the preceding example illustrates, tend to elicit only one explicit mental model in which the antecedent and the consequent both hold and one mental model with no explicit content. The latter model corresponds to the possibilities in which the antecedent of the conditional is false (see, e.g., Johnson-Laird & Byrne, 1991). Disjunctions, however, elicit multiple models with explicit content, one for each possibility, and one of them is likely to suggest a counterexample to the putative conclusion.

EXPERIMENT 2

Invalid conclusions are of two sorts: those that are inconsistent with the premises and those that are consistent with the premises but do not follow necessarily from them. In Experiment 1, we examined only invalid conclusions that were consistent with the premises, but in the present experiment, we examined both sorts of invalid conclusions. As we explained in the introduction, the model theory predicts that it should be easier to reject as invalid those conclusions that are inconsistent with the premises than those conclusions that are consistent with the premises. For those inferences in which the conclusion is inconsistent with the premises, every model of the premises is a counterexample to the conclusion, whereas the danger with consistent conclusions is that individuals fail to find a counterexample. The theory also predicts a difference in the strategies used to reject the two sorts of conclusion. Reasoners can establish the invalidity of an inconsistent conclusion by detecting the contradiction between the premises and the conclusion. That is, they can determine that there is no model in which the premises and the conclusion are true (Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000), or they could use a formal procedure to prove that the premises contradict the conclusion (Braine & O'Brien, 1991). However, this strategy, whether implemented by models or rules, cannot be used to establish the invalidity of conclusions that are consistent with the premises; these conclusions call for the construction of counterexamples. The model theory accordingly predicts a more frequent use of counterexamples for such problems. But since the use of counterexamples is error prone (i.e., it is possible to miss a counterexample and, therefore, wrongly accept an invalid conclusion), it follows that participants should be more accurate in noticing the invalidity of inconsistent problems. In Experiment 2, we examined these predictions of the model theory.

Method

Materials and Procedure. In Experiment 2, we tested a new set of 28 participants from the same population as before. They acted as their own controls and evaluated 10 pairs of premises presented once with an invalid conclusion consistent with the premises and once with an invalid conclusion inconsistent with the premises, but in separate blocks of trials and with different contents. The forms of the problems are summarized in Table 3. There were also six filler problems with valid conclusions to control for response bias. The design, procedure, and materials were as similar as possible to those in Experiment 1 and, again, called for the participants to write justifications for their evaluations of the given conclusions.

Results and Discussion

Table 3 presents the percentages of counterexamples in which the participants responded correctly to each of the problems with the two sorts of invalid conclusions. Table 4 summarizes the percentages with which the three main sorts of strategy occurred for the two sorts of invalid conclusion. The results corroborated both predictions of the model theory. First, the participants were more accurate in recognizing that a conclusion was invalid when it was inconsistent with the premises (92%) than when it was consistent with the premises (74%; Wilcoxon test, $z = 3.2$, $p < .001$, by participants; $z = 2.93$, $p < .01$, by items). Second, as Table 4 shows, the participants used counterexamples to refute invalid conclusions consistent with the premises more often than they used counterexamples of invalid conclusions inconsistent with the premises (Wilcoxon test, $z = 1.95$, $p < .05$, by participants; $z = 2.8$, $p < .005$, by items). When the conclusion was inconsistent with the premises, the participants tended instead to point out the contradiction. Table 2 shows examples of the justifications. To test the reliability of the classification, two judges independently classified justifications from 7 participants (108 correct justifications) into the three main categories: the use of a counterexample, the detection of a contradiction between the premises and the conclusion, and the remain-

Table 3
The Forms of Problem in Experiment 2, With Both Inconsistent and Consistent Invalid Conclusions, and the Percentages of Counterexamples Used in Their Correct Evaluations

Problem	Form of Premises	Inconsistent Conclusion	Percentage of Counterexamples	Consistent Conclusion	Percentage of Counterexamples
1	A or else B B or else C	∴ not A and C	50	∴ not A and not C	63
2	A or else B B or C	∴ A and not C	46	∴ not A and not C	60
3	A or else B If B then C	∴ not A and not C	33	∴ A and not C	74
4	A or B B or else C	∴ not A and C	46	∴ A and not C	67
5	A or B B and C	∴ A and not C	17	∴ not A and C	43
6	A or B If B then C	∴ not A and not C	36	∴ not A and C	37
7	A and B B or C	∴ not A and C	20	∴A and not C	42
8	If A then B B or else C	∴ A and C	16	∴ A and not C	54
9	If A then B B and C	∴ A and not C	4	∴ not A and C	19
10	If A then B If B then C	∴ A and not C	17	∴ not A and C	53
	Overall		29		51

ing miscellaneous strategies. They agreed on 91% of the evaluations (Kendall's tau = .5, $z = 8.75$, $p < .001$). Of the 28 participants, 26 used counterexamples at least once (Binomial test, assuming a prior probability of .5, $p < .00001$). Likewise, for 16 of the 20 problems, counterexamples were the most frequently used strategy (Binomial test, $p < .01$). The proportion of correct evaluations on which participants used counterexamples correlated with the total number of correct evaluations, and we assessed the correlation in the same way as in Experiment 1 (Pearson's $r = .58$, $p < .001$). Indeed, of the 2 participants who never used counterexamples, 1 of them erroneously evaluated all the invalid–consistent conclusions. This participant was able, however, to detect inconsistencies between a conclusion and the premises. The other participant who never used counterexamples relied on the strategy of detecting that a necessary piece of information was missing from the premises. For example, given a problem of the form

A or else B, but not both.

B or else C, but not both.

Therefore, not-A and C.

the participant noted that there was nothing in the premises that necessarily implied *not-A*.

EXPERIMENT 3

Certain theories of reasoning make no provision for the use of counterexamples whatsoever, and so the critical issue is whether in any circumstances naive individuals use them. The two previous experiments showed that when individuals have to write justifications for their evaluations of given conclusions, they do indeed rely on counterexamples. But the need to write justifications could elicit strategies different from those that they would otherwise use. Perhaps naive reasoners would not use counterexamples when they merely evaluated conclusions without the need to justify their responses. In the previous experiment, as the model theory predicted, the participants were less accurate in refuting invalid conclusions that were consistent with the premises than in refuting invalid conclusions that were inconsistent with them. In the latter case, there is less chance of missing a counterexample to the conclusion. One of the goals of the present experiment was to evaluate whether this difference in accuracy would hold for participants who did not write justifications. If there were no difference between the two sorts of invalid conclusion under such a condition, the model theory would be at risk. Another sign of the use of different strategies in this condition would be a different pattern of latencies. The purpose of

Table 4
The Percentages of Use of the Three Main Sorts of Strategies for Consistent and Inconsistent Invalid Conclusions in Experiment 2

Conclusion	Counterexamples	Contradictions	Miscellaneous Strategies
Consistent	51	0	49
Inconsistent	29	53	18

the present experiment was to examine these issues. It compared accuracy and latency for participants who wrote justifications for their evaluations with those for participants who did not write justifications.

Method

Materials and Procedure. We tested 20 new participants from the same population as before. The materials were identical to those in Experiment 2. The problems were presented individually on a computer, and the participants were told that their task was to determine whether the conclusion necessarily followed from the premises. They were told to press the "J" key if the conclusion necessarily followed from the premises and the "F" key if it did not. The presentation of the problems was self-paced, and the participants were told to consider each problem for just the time needed to solve it. This experimental procedure was for the participants who only evaluated conclusions (the *no-justification* group henceforth). The participants in a second group were asked in addition to write why they thought that the conclusion did or did not follow from the premises (the *justification* group). For this group, each trial was followed by an additional screen in which the original problem was presented above a text box, in which the participants typed their justifications. The latency measure for both groups was from the time of the initial presentation of a problem to the participants' decision about its conclusion.

Results and Discussion

Table 5 presents the accuracy results and the response latencies for both groups of participants. Response latencies were collapsed across correct and incorrect evaluations. There was no reliable difference between the groups in accuracy. Accuracy was 74% in the justification condition and 72% in the no-justification condition. To maximize the power of the analysis, we carried out an analysis of variance with justification condition as a between-subjects variable and the sort of conclusion, consistent or inconsistent, as a repeated measure. It revealed no main effect of justification and no significant interactions with this variable (Fs < 1). However, as the model theory predicts, the participants were more accurate in recognizing that inconsistent conclusions did not follow from the premises (86% correct) than in recognizing that consistent conclusions did not follow from the premises [60% correct; $F(1,18) = 27.95$, $p < .001$]. The difference was reliable for both the justification and the no-justification groups [$t(10) = 4.97$, $p < .001$, and $t(10) = 3.44$, $p < .01$, respectively]. Although there was a tendency for the participants to take longer to respond in the justification group (17.7 sec) than in the no-justification group (15.3 sec), the difference was not reliable, and neither was the interaction (Fs < 1). The evaluation task might take longer when participants know that they have to write a justification for their evaluation, and so it is slightly surprising that the difference was not reliable.

For the 10 participants who wrote justifications, we evaluated the tendency to use counterexamples for problems with consistent and inconsistent conclusions. As before, we categorized a justification as a counterexample if it contained an explicit mention of a possibility that was consistent with the premises but not with the conclusion. The participants tended to use counterexamples more often in refuting conclusions that were consistent with the premises (34% of the trials) than in refuting conclusions that were inconsistent with the premises (8% of the trials, Wilcoxon test, $z = 2.41$, $p < .05$, by participants; $z = 1.98$, $p < .05$, by materials). There was once more a positive, although unreliable, correlation between the proportion of counterexamples and overall accuracy, which we assessed in the same way as in the previous experiments (Pearson's $r = .46$, $p = .18$).

The similarity in the pattern of accuracy and latency suggests that the participants in the two groups were not using grossly different strategies. Whether or not they had to write a justification, it remained more difficult to reject as invalid a conclusion that was consistent with the premises than to reject a conclusion that was inconsistent with the premises. This difference is unlikely to be attributable to a failure to understand the instructions. They made clear that a conclusion must follow "necessarily" from the premises to elicit a positive evaluation, and the participants asserted that they understood this point. Hence, as far as we can tell, individuals are able to use counterexamples to refute invalid conclusions whether or not they have to justify their evaluations.

Table 5
The Percentages of Accurate Evaluations and the Overall Latencies (in Seconds) for Evaluations in Experiment 3 for the Justification and No-Justification Groups

	Inconsistent and Invalid Conclusion		Consistent but Invalid Conclusion	
Group	Accuracy	Latency	Accuracy	Latency
Justification	85	18.43	63	17.04
No justification	87	14.93	56	15.60

GENERAL DISCUSSION

The use of counterexamples to establish invalidity is theoretically neutral between the use of models and the use of formal rules. Current psychological theories of both sorts, however, have downplayed counterexamples (see, e.g., Polk & Newell, 1995; Rips, 1994). In contrast, the mental model theory predicts that individuals should be able to use counterexamples to refute invalid conclusions (e.g., Johnson-Laird & Byrne, 1991). Experiment 1 corroborated this prediction. It showed that individuals develop various strategies of refutation. And the most frequent strategy was to envisage models of the premises in which the conclusion is false and to describe such models as their reason for rejecting a conclusion consistent with the premises. These written justifications made clear that the participants were aware that such possibilities serve as counterexamples to conclusions. Experiment 2 corroborated a further prediction of the model theory: Individuals were more likely to use a counterexample to refute a conclusion consistent with the premises than to use one to refute a conclusion inconsistent with the premises.

Our chief interest was whether individuals can use counterexamples in any circumstances. But to what extent are they likely to do so when they do not have to write justifications for their responses? We have several grounds for supposing that they can still use counterexamples. Experiment 3 compared performance between two groups of participants: One group wrote justifications, and the other group did not. This manipulation had no reliable effect on the accuracy or latency of their responses. Moreover, both groups corroborated the model theory's prediction that invalidity was harder to detect with conclusions consistent with the premises than with conclusions inconsistent with them. Other evidence comes from studies of reasoning based on quantifiers. These have shown that reasoners spontaneously construct external models as counterexamples (Bucciarelli & Johnson-Laird, 1999) and that they spontaneously draw diagrams that serve as counterexamples (Neth & Johnson-Laird, 1999). Preliminary brain-imaging studies have suggested that the frontal pole in the right frontal cortex may become activated in reasoning only when individuals search for counterexamples (Kroger, Cohen, & Johnson-Laird, 2002). If this finding proves to be robust, it may be possible to use activation of the frontal pole as a sign that reasoners are using counterexamples.

Individuals from other cultures or subcultures may fail to use counterexamples (cf. Peng & Nisbett, 1999). The propensity to use them varies in all of our three experiments, and a few participants did not use them at all. Several factors appear to give rise to this variation. One likely factor is intellectual ability: The use of counterexamples correlated with correct evaluations of conclusions. Granted that accurate performance in reasoning depends on intellectual ability (see Stanovich, 1999), the failure to use counterexamples may be a consequence of lack of ability or perhaps a limited processing capacity in working memory (see, e.g., Barrouillet & Leças, 1999). Another factor in the use of counterexamples is whether an invalid conclusion is consistent or inconsistent with the premises (see Experiments 2 and 3). The participants often refuted inconsistent conclusions by detecting the contradiction between the conclusions and the premises; but consistent conclusions cannot be refuted in this way, and so the participants were more likely to envisage a counterexample. Still other factors are likely to underlie the use of counterexamples. When reasoners draw their own conclusions, they may be able to keep in mind all the possibilities compatible with the premises and, therefore, have no need to search for counterexamples. In general, the best recipe to elicit a search for counterexamples seems to be to ask highly intelligent individuals to evaluate a given conclusion, to use premises that elicit multiple models of possibilities, and to ensure that the conclusion holds in at least one of the models.

Theories of reasoning that aim for a complete account of human competence must allow for the role of counterexamples. It is contrary to theories that postulate a single deterministic strategy in which invalidity is established solely by other means (cf. Rips, 1994). But it does not rule out theories of reasoning based on formal rules of inference. Formal theories can mimic the operation of constructing counterexamples. The chief difficulty is to cope with the refutation of conclusions consistent with the premises—for example, Problem 3 in Table 3, which is of the form

A or else B.

If B then C.

Therefore, A and not C.

You might suppose that invalidity could be demonstrated directly by the formal rule of conditional proof—for example,

Suppose A.

Therefore, not B (from the first premise).

At this point, no further valid inference is possible. A supposition of *not C* even yields a derivation of *A* and, hence, the conditional conclusion *If not C then A*. Hence, the most that can be done with formal rules of the sort currently postulated in psychological theories is to fail to find a proof of invalid conclusions consistent with the premises. To cope with such invalid conclusions, the best strategy when formal rules are used is to adopt a procedure based on the tree method (see, e.g., Jeffrey, 1981). In this procedure, formal rules are used to make an exhaustive search for counterexamples. The search begins with a list of the premises and the *negation* of the putative conclusion. These statements can be true only if there is a counterexample. And so the rules of inference are formulated in a way that allows for a systematic search of the possibilities. No proponent of formal rules in psychology, however, has seriously proposed such an account: Naive individuals do not appear to start reasoning by *negating* the conclusion to be evaluated, and none of the participants in our experiments opted to prove the negation of a conclusion.

The model theory postulates that what lies at the heart of human rationality is a grasp of the fundamental *semantic* principle of validity (Beth, 1971): An inference is valid if its conclusion holds in all possibilities consistent with the premises. One application of this principle to the evaluation of invalidity is to construct counterexamples, and our results suggest that naive individuals can grasp their force.

REFERENCES

Barres, P., & Johnson-Laird, P. N. (2003). On imagining what is true (and what is false). *Thinking & Reasoning*, **9**, 1-42.

Barrouillet, P., & Leças, J.-F. (1999). Mental models in conditional reasoning and working memory. *Thinking & Reasoning*, **5**, 289-302.

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

Bell, V., & Johnson-Laird, P. N. (1998). A model theory of modal reasoning. *Cognitive Science*, **22**, 25-51.

Beth, E. W. (1971). *Aspects of modern logic*. Dordrecht: Reidel.

Braine, M. D. S., & O'Brien, D. P. (1991). A theory of if: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review*, **98**, 182-203.

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

Evans, J. St. B. T., Handley, S. J., Harper, C. N. J., & Johnson-Laird, P. N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 1495-1513.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction.* Mahwah, NJ: Erlbaum.

Holyoak, K. J., & Glass, A. (1975). The role of contradictions and counterexamples in the rejection of false sentences. *Journal of Verbal Learning & Verbal Behavior*, **14**, 215-239.

Jeffrey, R. (1981). *Formal logic: Its scope and limits* (2nd ed.). New York: McGraw-Hill.

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

Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, **109**, 646-678.

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

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J.-P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, **106**, 62-88.

Kroger, J. K., Cohen, J. D., & Johnson-Laird, P. N. (2002). *A double dissociation between logic and mathematics.* Manuscript submitted for publication.

Neth, H., & Johnson-Laird, P. N. (1999). The search for counterexamples in human reasoning. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society* (p. 806). Mahwah, NJ: Erlbaum.

Newstead, S. E., Handley, S. J., & Buck, E. (1999). Falsifying mental models: Testing the predictions of theories of syllogistic reasoning. *Memory & Cognition*, **27**, 344-354.

Newstead, S. E., Thompson, V. A., & Handley, S. J. (2002). Generating alternatives: A key component in human reasoning? *Memory & Cognition*, **30**, 129-137.

Peng, K., & Nisbett, R. E. (1999). Culture, dialectics, and reasoning about contradiction. *American Psychologist*, **54**, 741-754.

Polk, T., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, **102**, 533-566.

Quine, W. V. O. (1974). *Methods of logic* (3rd ed.). London: Routledge.

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

Roberts, M. J. (in press). Falsification and mental models: It depends on the task. In W. Schaeken, A. Vandierendonck, W. Schroyens, & G. d'Ydewalle (Eds.), *The mental models theory of reasoning: Refinement and extensions*. Mahwah, NJ: Erlbaum.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning.* Mahwah, NJ: Erlbaum.

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

(Manuscript received October 30, 2002;
revision accepted for publication May 9, 2003.)