Hempel's Paradox and Wason's Selection Task: Logical and Psychological Puzzles of Confirmation

Raymond S. Nickerson

Tufts University, Medford, USA

Hempel's paradox of the ravens has to do with the question of what constitutes confirmation from a logical point of view; Wason's selection task has been used extensively to investigate how people go about attempting to confirm or disconfirm conditional claims. This paper presents an argument that the paradox is resolved, and that people's typical performance in the selection task can be explained, by consideration of what constitutes an effective strategy for seeking evidence of the tenability of universal or conditional claims in everyday life.

INTRODUCTION

Confirmation has more than one connotation, as they term is used in everyday language. Sometimes it is intended to be synonymous with *proof* or *verification* and to indicate establishment of the truth of something with certainty; and sometimes it conveys the idea of *support* or *corroboration* . An observation is said to confirm a claim or hypothesis in the latter sense if it justifiably increases its credibility, even though it does not establish truth beyond doubt.

These two connotations have been referred to, respectively, as the absolute and the incremental senses of the term (Poundstone, 1990). The focus in this paper is on confirmation in the second, incremental, sense: confirmatory evidence is taken to be evidence that provides support to a claim or hypothesis. This connotation admits the possibility of finding confirmatory evidence for a claim or hypothesis that, with complete information, would be recognised to be false.

Underlying the notion of incremental confirmation is the idea that a case of a hypothesis supports that hypothesis—that the observation of a black raven, for example, supports the hypothesis that all ravens are black. This idea has been challenged. Good (1967) gives examples of situations in which a case of a

Requests for reprints should be sent to Raymond S. Nickerson, Department of Psychology, Tufts University, Medford, MA, 02155, USA.

I am indebted to David Green, Kenneth Manktelow, David Over, and Rosemary Stevenson for pointing out to me several of the more recent studies cited and for other helpful comments on a draft of this article.

hypothesis can be disconfirmatory with respect to that hypothesis. Without denying the validity of Good's argument or the legitimacy of his examples, I think the problem can be ignored for present purposes, on the grounds that situations of the sort he describes are not characteristic of those typically encountered in everyday life. It is assumed here that in the vast majority of real-life situations calling for evaluation of hypotheses, a case of a hypothesis *does* support the hypothesis.

The nature of confirmation is the subject of an old and continuing debate (Salmon, 1973). This paper is concerned with two well-known aspects of the debate—Hempel's paradox and Wason's selection task. Both of these topics have received a great deal of attention, the former primarily from philosophers and the latter from psychologists. The purpose of this paper is to consider what light, if any, either topic sheds on the other.

THE PARADOX OF THE RAVENS

In 1945, the philosopher Carl Hempel presented a problem of reasoning relating to confirmation which evoked discussion that continues to the present. Consider the assertion "All ravens are black". On the assumption that a case of a hypothesis supports that hypothesis, observations of black ravens should strengthen our confidence in the truth of this assertion.

Hempel pointed out that the assertions "All ravens are black" and "All nonblack things are nonravens" are logically equivalent, each being the contrapositive of the other. Any evidence that strengthens belief that one of them is true should strengthen belief that the other (equivalent) one is true as well. By the rule that a case of a hypothesis supports that hypothesis, the observation of a white shoe should increase our confidence in the truth of the second assertion. And given that the two statements are equivalent, we have no choice, Hempel argues, but to take the same observation as confirmatory evidence also for the claim that all ravens are black.

But who would do so? Most of us would consider the observation of a white shoe to have little, if anything, to do with the claim that all ravens are black. We would be especially reluctant to take the observation of a white shoe as confirmation of this claim upon realising that the logic that leads to the conclusion that we should do so leads also to the conclusion that we should take the same observation as confirmation of the contradictory claim that all ravens are red, because the contrapositive of the latter claim is that all nonred things are nonravens.

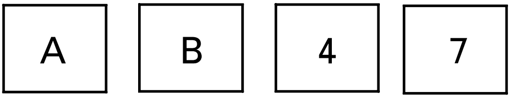
WASON'S SELECTION TASK

About 20 years after Hempel first described his paradox, Peter Wason (1966, 1968) invented a task for use in the study of conditional reasoning, and reported the results of an initial experiment that stimulated a host of subsequent studies and much discussion and theorising about the capabilities and limitations of humans as intuitive logicians. Wason's (1966, p.145) description of the task was as follows:

> The subjects (students) were presented with an array of cards and told that every card had a letter on one side and a number on the other side, and that either would be face upwards. They were then instructed to decide which cards they would *need* to turn over in order to determine whether the experimenter was lying in uttering the following statement: *If a card has a vowel on one side then it has an even number on the other side.*

Wason considered the correct response to be selection of cards displaying either a vowel or an odd number, because only the discovery of a card with this combination would prove the statement false. He found the most frequent response to be selection of cards displaying a vowel and those displaying an even number; no-one selected cards showing a consonant and only a small minority selected those showing an odd number.

Although it is not clear from this description whether the array Wason originally used contained only four cards, this has been the case in many subsequent experiments. The following array is typical:



Given this set of cards, selection of those showing *A* and *7* has generally been considered the correct answer, because either of these cards could have a symbol on the back that would show the claim of interest to be false; the card showing the *B* and the one showing the *4* have been considered irrelevant, because whatever either has on its other side is consistent with the claim.

The typical finding of experiments with this task bears out Wason's original result: a small minority of participants select the two cards showing the vowel and the odd number; most select the cards showing the vowel and the even number, or only the one showing the vowel. Essentially the same finding has been obtained by numerous investigators with one or another variant of the selection task; the results of experimentation with this task have been reviewed many times (Cosmides, 1989; Evans, 1982; Evans, Newstead, & Byrne, 1993; Tweney & Doherty, 1983).

In most of the remainder of this article, I will use the notation "If *P* then *Q*" to discuss the selection task in general terms. Although as applied to Wason's original version of the task *P* corresponds to "a card has vowel on one side", and *Q* is "it has an even number on the other", relative to other versions of the task the symbols may have other referents; given, for example, the conditional

"If a letter is sealed then it has a 50 lira stamp on it", *P* represents "a letter is sealed", and *Q* "it has a 50 lira stamp on it".

A Critical Ambiguity

It has been claimed that Wason's selection task has generated more psychological research than any other single experimental paradigm. Among the attractions of this paradigm are its simplicity and the consistency with which it has yielded the particular result that people typically fail to see the relevance of the *~Q* option—the odd-number card in Wason's original experiment or its equivalent in other paradigms—which has been taken by many as evidence that people generally have difficulty thinking in accordance with the rules of conditional logic. The apparent simplicity of the task may be illusory, to an extent; there is an ambiguity in some statements of it that did not, for a long time, receive the attention it deserves.

Consider again the case in which one is shown four cards, and asked to say which of the cards must be turned over to determine the truth or falsity of the assertion "If a card has a vowel on one side, it has an even number on the other". In the absence of further qualifications, there are at least two ways to interpret the question that is being asked: (1) which *of the four cards in view* must be turned over in order to determine the truth or falsity of the assertion *with respect to those four cards*, and (2) which of the four *types* of cards represented should be turned over in order to determine the plausibility of the assertion *in general*, or at least with respect to a set of cards larger than the four showing, say a large deck from which the four were drawn. Sometimes the instructions have made it clear that the first of these interpretations is the intended one, but often they have not.

When it is clear that one's task is to say which of four visible cards must be turned over in order to determine whether a particular conditional of the form "If *P* then *Q*" is true of *those four card*s, there can be no question that the only correct answer is the card showing *P* and the one showing *~Q*. When the task can be interpreted as that of obtaining evidence regarding the truth or falsity of the assertion with respect to a larger set of cards for which the four in view are proxies, plausible rationales can be presented to justify other selections, depending on the specifics of the properties represented by *P* and *Q* and what is known or assumed about their prevalence in the population of interest.

In what follows, the *second* interpretation of the selection task is intended except when the context clearly indicates otherwise. This interpretation makes the selection task more representative, I think, of conditional reasoning of the sort that one is called upon to do in everyday life; seldom are we faced with the problem of determining the truth or falsity of a conditional claim involving only four entities, all of which are immediately available for inspection.

RELATIONSHIP BETWEEN HEMPEL'S PARADOX AND WASON'S SELECTION TASK

The assertion of interest in Hempel's paradox is stated in categorical terms, "All *P* (ravens) are *Q* (black)"; that in Wason's task is stated as a conditional, "If *P* (a card has a vowel on one side) then *Q* (it has an even number on the other)". But this difference is superficial. The claim regarding ravens could as well be stated as "If *P* (X is a raven), then *Q* (X is black)", and Wason's task could be stated as that of determining the truth or falsity of the categorical assertion "All *P* (cards with a vowel on one side) are *Q* (cards with an even number on the other)". In what follows, I shall most often use the conditional form, but will use the categorical form when it seems more appropriate to the context.

Logicians distinguish between two basic forms of the hypothetical syllogism: *modus ponens* and *modus tollens*. Letting ~ represent negation, *modus ponens* is:

If *P* then *Q*
P
Therefore *Q*

and *modus tollens*:

If *P* then *Q*
~*Q*
Therefore ~*P*.

What is required to demonstrate "If *P* then *Q*" to be *false* is the observation of something that is both *P* and ~*Q*. If something that is *P* and ~*Q* exists, it is a member both of the set defined by *P* and of the set defined by ~*Q*. One might say that *modus ponens* focuses the search for such a falsifying case on the set defined by *P* and that *modus tollens* focuses the search on the set defined by ~*Q*. Because in Wason's original task, *P* is "a card has a vowel on one side" and *Q* is "it has an even number on the other", turning the card that shows a vowel to see if it has an even number on the back is applying the *modus ponens* implication to the task, whereas turning over the card showing the odd number (~*Q*) to be sure that it is accompanied by the consonant (~*P*) is applying the *modus tollens* form.

The aspect of people's performance of Wason's task that has attracted the greatest attention is the common failure to select the card showing an odd number (~*Q*). This is often described as a failure to apply the *modus tollens* form of argument. The idea that people find it less natural to use *modus tollens* than *modus ponens* has considerable experimental support (Evans et al., 1993;

Rips & Marcus, 1977; Rumain, Connell, & Braine, 1983; Wason & Johnson-Laird, 1972).

However, just as "All nonblack things are nonravens" is the contrapositive of "All ravens are black", so "If ~*Q* then ~*P*" is the contrapositive of—and thus logically equivalent to—"If *P* then *Q*". If one recognises this equivalence, one can identify the card showing an odd number by applying *modus ponens* to the first of the latter two assertions. So an alternative way of describing the failure to specify ~*Q* in the selection task is as a failure to recognise the equivalence of an assertion and its contrapositive.

This language is essentially the same as that which has been used to describe Hempel's paradox. Is there some relationship between Wason's card showing the odd number and Hempel's white shoe? Is the tendency of people to overlook the relevance of the former related to the difficulty we have in seeing why the observation of a white shoe should increase our confidence that all ravens are black?

IMPLICATIONS OF "IF P THEN Q"

As it is used in everyday language, "if" is ambiguous; normally the intended meaning is disambiguated by the context in which it occurs. In logic an assertion of the form "If *P* then *Q*" is defined by its truth function. As shown in Table 1, the assertion is said to be true when both *P* and *Q* are true, when *P* is false and *Q* is true, and when both *P* and *Q* are false; it is said to be false only when *P* is true and *Q* is false. An alternative way of showing the conditional is as in Table 2, where the cell entries indicate which combinations of *P* or ~*P* and *Q* or ~*Q* are consistent with the assertion, "If *P* then *Q*." (Conventionally, *P* is called the antecedent, and *Q* the consequent.)

I like the latter way of representing the implications of the conditional, because it makes more immediately obvious than does the truth-function representation the fact that knowing the antecedent to be true (*P*) permits one to make an inference about the consequent, and knowing the consequent to be false (~*Q*) permits one to make an inference about the antecedent, whereas knowing the antecedent to be false (~*P*) does not permit an inference about the consequent, and knowing the consequent to be true (*Q*) does not permit an

TABLE 1
The Truth Function of "If P then Q"

P	*Q*	*If P then Q*
T	T	T
T	F	F
F	T	T
F	F	T

TABLE 2
An Alternative Way of Showing the Conditional

	Consequent	
Antecedent	*Q*	*~Q*
P	Yes	No
~P	Yes	Yes

Each cell entry indicates whether the combination specified by the associated row and column headings is consistent with the assertion "If *P* then *Q*".

inference about the antecedent. The first row of the Table shows that if *P* is true, *Q* must be true, and the second column shows that if *Q* is false, *P* must be false; so if one knows the antecedent to be true, one knows the consequent must be true (*modus ponens*), and if one knows the consequent to be false, one knows the antecedent must be false (*modus tollens*). But because a false *P* can be associated with either a true or false *Q* (second row) and a true *Q* can be associated with either a true or false *P* (first column), nothing can be inferred from a knowledge of either of these states.

Although Table 2 does a better job than Table 1 of explicating the implications of the conditional, it still obscures an interesting fact. If we assume that none of the sets, *P*, *~P*, *Q*, or *~Q*, is empty, the relationship "If *P* then *Q*" *requires* that the combinations *PQ* and *~P~Q* exist, and it disallows the combination *P~Q*. And it allows, but does not require, that the combination *~PQ* exists. (Here and in what follows, I use juxtaposition to represent conjunction, so *PQ* represents something that is both *P* and *Q*, e.g. a black raven.) The situation is shown in Table 3.

That the *~PQ* combination bears a relationship to the conditional which differs from that of the other two combinations that are also consistent with it, follows from the fact that "If *P* then *Q*" is consistent both with the situation in which *P* is a proper subset of *Q* (as dogs are a proper subset of mammals) and

TABLE 3
Combinations Required, Allowed, and Disallowed

	Consequent	
Antecedent	*Q*	*~Q*
P	Required	Disallowed
~P	Allowed	Required

Each cell entry indicates whether the combinations specified by the associated row and column headings is required, allowed, or disallowed, assuming the assertion "If *P* then *Q*" is true and none of the sets, *P, ~P, Q*, and *~Q* is empty.

with that in which *P* and *Q* are the same set (as are spouses and marriage partners). In the latter case, in which *P* and *Q* are the same set, *~PQ* is empty.

JUDGING THE TRUTH OR FALSITY OF "IF P THEN Q"

The just-preceding comments centre on the implications of the assumption that "If *P* then *Q*" is true. But suppose the truth of the assertion is in question, as it is in the selection task, as first described by Wason. Here I want to consider the subject in general terms: the question is, how should one judge the tenability of a hypothesis of the form "If *P* then *Q*"?

It should be clear that the (only) way to show the hypothesis to be false, if it is false, is to determine that the combination *P~O* exists, and that this suffices to show *conclusively* that the hypothesis is false. Conversely, the only way to show, conclusively, that the hypothesis is true, if it is true, is to demonstrate that this combination does *not* exist. In many—perhaps most—real-world situations of interest, it is not possible to do the exhaustive testing necessary to satisfy the latter requirement, so typically general assertions are not said to have been demonstrated to be true in the absolute sense of it having been determined by exhaustive examination that no falsifying evidence exists. But, in the absence of counterindications, we do—in accordance with the case principle mentioned at the outset—accept as inductive evidence of a general assertion the observation of cases of the generalisation that is asserted.

To be specific: Because both *PQ* and—by the rule of the contrapositive—*~P~Q* are clearly cases of the generalisation, "If *P* then *Q*", we would consider the observation of either of these combinations as confirmatory of the hypothesis. But what about *~PQ*, which is allowed by the hypothesis but not required of it? The general question is: what, assuming we accept the notion that a hypothesis is confirmed by a case of that hypothesis, should be considered a "case of a hypothesis"? Should we think of a case of a hypothesis as something that the hypothesis *allows*, or only as something the hypothesis implies *must* be true? There can be little doubt that, at least as conditional reasoning is done in science, an observation that is predicted by a hypothesis is seen as more confirmatory of the hypothesis than one that is merely allowed—not prohibited—by it.

Hempel (1945) argued that the *~PQ* combination should be considered confirmatory because "All ravens are black" is equivalent to "A thing is either not a raven or is black, or both"—in terms of the conditional, "If *P* then *Q*" is equivalent to "Either *~P* or *Q* or both". It follows, Hempel argued, that "All ravens are black" is confirmed not only by a white shoe but by a black one as well, or indeed by anything, of any colour, that is not a raven. Hempel asserts (1945, p.14): "By virtue of the equivalence condition, we have therefore to consider confirming for S_1 {'All ravens are black'}, any object which is either no raven or also black (in other words: any object which is no raven at all, or a black raven)".

As the reader can easily verify, the truth function of the inclusive disjunctive "*~P* or *Q* or both" is indeed the same as that of "If *P* then *Q*". But if one accepts the foregoing argument that the existence of *~PQ* is not required by the truth of "If *P* then *Q*", as is that of *PQ* and *~P~Q*, then one might have reservations about the inclusiveness of Hempel's claim here. We shall see presently that under certain plausible assumptions, a Bayesian view of the situation would interpret *~PQ* as disconfirmatory.

In sum, we may say that the observation of *P~Q* is, by deductive logic, conclusively disconfirmatory, and the observation of either *PQ* or *~P~Q* is, by induction, inconclusively confirmatory. If one accepts Hempel's claim, the observation of *~PQ* is also inconclusively confirmatory, but, for the moment, let us withhold judgment on the status of this combination. So, given the question of whether a generalisation of the sort "If *P* then *Q*" is true, if I observe an instance of *P~Q*, I know it to be false; if I observe an instance of either *PQ* or *~P~Q* (while *not* observing any instances of *P~Q*), I increase my confidence somewhat that it is true. If I observe *~PQ*, it is not yet clear what I should do.

These comments have focused on antecedent–consequent *combinations*, and, in particular, on the confirmatory or disconfirmatory character of possible combinations of *P* or *~P* and *Q* or *~Q* with respect to the conditional assertion "If *P* then *Q*". In the selection task, as originally conceived, one does not initially see a *combination* of *P* or *~P* and *Q* or *~Q*; one sees each of four cards with the elements, *P, ~P, Q,* and *~Q*. But one knows that each of the cards showing *P* or *~P* has *Q* or *~Q* on its hidden side, and that each of the cards showing *Q* or *~Q* has *P* or *~P* on its hidden side.

All this being true, one should know that the conclusively disconfirmatory combination, *P~Q* can be found (if it exists) only by turning over either the card showing *P* or the one showing *~Q*. One should know, too, that the inconclusively confirmatory combination *PQ* might be found by turning over either the card showing *P* or the one showing *Q*, and that the inconclusively confirmatory combination *~P~Q* might be found by turning over the card showing *~P* or the one showing *~Q*. It would seem, according to this line of reasoning, that *any* of the cards *could* yield confirmatory information. Why, if that is the case, is not the correct response in the selection task to turn over all the cards? The answer is obvious for the case in which the task is to specify which of four cards must be turned over in order to determine the truth or falsity of the assertion with *respect to those four cards*, but recall that we are now focusing on the case in which the task is interpreted as that of specifying which of the four *types* of cards represented should be turned over in order to determine the truth or falsity of the assertion *in general*.

We will return to the question of whether selection of all four cards should be considered the correct answer in the general case. Suffice it to emphasise at this point that, although any of the cards could yield inconclusively confirmatory information, conclusively disconfirmatory information can be produced by only two of them, *P* and ~*Q*. The fact that performers of the selection task commonly select the card representing *Q* (the even number, in the case of our example) as well as the one representing *P* (the vowel) suggests that, given the task of assessing the credibility of a conditional assertion, people are more likely to look for confirmatory than for disconfirmatory information. The fact that they typically select the card representing *P* but not the one representing ~*Q* suggests that they are more likely to look for instances in which the consequent is present when the antecedent is known to be present, than to look for the absence of the antecedent when the consequent is known to be absent. In terms of Hempel's paradox, they are more likely to look for black ravens than for either white ravens or white shoes.

NICOD'S CRITERION OF CONFIRMATION

Hempel discussed a conception of confirmation put forward by Jean Nicod (1930), which purports to explain how a fact can affect the probability of a law of the form *P entails Q*. According to Nicod (1930, p.219):

> If this fact consists of the presence of Q in a case of P, it is favourable to the law *"P entails Q";* on the contrary, if it consists of the absence of Q in a case of P, it is unfavourable to this law. It is conceivable that we have here the only two direct modes in which a fact can influence the probability of a law . . . Thus, the entire influence of particular truths or facts on the probability of universal propositions or laws would operate by means of these two elementary relations which we shall call *confirmation* and *invalidation.* {I have substituted *P* and *Q*, where Nicod used *A* and *B*, respectively.}

Hempel criticised Nicod's criterion on two grounds: (a) that it is applicable only to hypotheses of universal conditional form and not, for instance, to existential hypotheses (There is life outside the solar system), and (b) that it does not necessarily recognise an observation that is confirmatory of a given statement as confirmatory also of any logically equivalent statement—to wit, it does not recognise the observation of a nonblack nonraven, which confirms the statement that all nonblack things are nonravens, as confirming of the logically equivalent claim that all ravens are black.

The second shortcoming means "that Nicod's criterion makes confirmation depend not only on the content of the hypothesis, but also on its formulation" (Hempel, 1945, p.ll). This violates the *equivalence condition*, according to which, "Whatever confirms (disconfirms) one of two equivalent sentences, also confirms (disconfirms) the other" (Hempel, 1945, p.12). Hempel argues (1945, p.13) that "the equivalence condition has to be regarded as a necessary condition for the adequacy of any definition of confirmation", which is to say that if the observation of a white shoe is to be taken as confirming of the claim that all nonblack things are nonravens, it must also be considered confirmatory of the claim that all ravens are black.

HEMPEL'S RESOLUTION OF HIS PARADOX

Hempel (1945, p.18) argued that the paradox of the ravens is a paradox in appearance only—a psychological illusion resulting from reliance on a misleading intuition:

> One source of misunderstanding is the view . . . that a hypothesis of the simple form "Every *P* is a *Q*", such as "All sodium salts burn yellow", asserts something about a certain limited class of objects only, namely, the class of all *P*s. This idea involves a confusion of logical and practical considerations: Our interest in the hypothesis may be focused upon its applicability to that particular class of objects, but the hypothesis nevertheless asserts something about, and indeed imposes restrictions upon, *all* objects (within the logical type of the variable occurring in the hypothesis, which in the case of our last illustration might be the class of all physical objects). Indeed, a hypothesis of the form "Every *P* is a *Q*" forbids the occurrence of any objects having the property *P* but lacking the property *Q*; i.e., it restricts all objects whatsoever to the class of those which either lack the property *P* or also have the property *Q*. Now every object either belongs to this class or falls outside it, and thus, every object—and not only the *P*s—either conforms to the hypothesis or violates it; there is no object which is not implicitly "referred to" by a hypothesis of this type. In particular, every object which either is no sodium salt or burns yellow conforms to, and thus "bears out" the hypothesis that all sodium salts burn yellow; every other object violates that hypothesis.

(By hypothesis, the class containing "every other object" is empty; that is, if the hypothesis is true, there are no objects that are sodium salts and fail to burn yellow.)

Hempel argued, too, that the illusion rests also in part on the fact that, as the paradox is presented here, the ~*Q* object is something that we already know is ~*P* (we know that a white shoe is not a raven), so we gain no information that is relevant to the hypothesis that every *P* is a *Q* by observing one. In fact, this points up an important difference between the situation considered by Hempel and the selection task. Observing an odd number on a card (~*Q*) in the selection task is not quite equivalent to observing a white shoe in the situation considered by Hempel. In the latter case, one knows that the object is white *and* that it is a shoe; in the former one knows only that one side of the card shows an odd number—before turning it over one does not know what is on the other side.

The raven-paradox equivalent of being shown the one side of a card displaying an odd number in the selection task would be to be given a box and told only that it contains a nonblack *object*. Even people who believe that observation of a white shoe should be considered irrelevant to the claim that all ravens are black might wish to open the box to satisfy themselves that the nonblack thing in it is not a raven. In any case, according to the logic of the contrapositive, upon opening a box known to contain a white object, one obtains information—learns something—that is confirmatory of the claim that all ravens are black if the opened box reveals a white shoe. More generally, if given the box and told nothing about its contents, one gets evidence that is *consistent* with the hypothesis if, upon opening it, one finds *anything* other than a nonblack raven . . . a black raven, a white shoe, a black shoe, or a confused logician.

THE PRINCIPLE OF FALSIFICATION AND SEARCH EFFICIENCY

It is a first principle of reasoning that universal claims about innumerable sets in the physical world (all ravens) cannot be proved—evidence can be marshalled that increases their tenability, but they cannot be verified with certainty. Sometimes they can be shown definitely to be false if they are false, but they cannot be shown definitely to be true if they are true.

From the Popperian perspective, the best way to marshall evidence for a universal statement is to fail to falsify it despite trying very hard to do so (Popper, 1959). What is needed to falsify the claim that all ravens are black is, of course, the observation of one nonblack raven. Everyone, even people who may disagree on the question of whether the observation of a black raven and that of a white shoe should or should not be considered equally confirmatory, will agree on this point. So, according to this view, the way to test the claim that all ravens are black is to try very hard to find (at least) one nonblack raven. A practical question that arises is that of how to conduct a search so as to minimise the effort required to find such a creature if one exists. Where does one look to maximise the chances of finding a counterexample to the claim if there is one to be found?

Most of us, I suspect, would go looking for ravens, noting, whenever we found one, whether or not it was black; we would be unlikely to go looking for nonblack objects, noting for each one that was found whether or not it was a raven. Hempel's argument notwithstanding, the second strategy would appear to be a singularly unproductive one. Wetherick (1993) appeals to this idea as an explanation of why people performing the selection task select *P* but not ~*Q*. If people interpret the selection task as "a laboratory model of the real-world task", as he suggests they do, their failure to select ~*Q* is just an analogue of the reasonable decision not to look for a nonblack raven by searching the universe of nonblack things. The assumption that there are many more nonblack things than ravens in the world is critical to this line of reasoning. If it were the case that ravens outnumbered nonblack things, then the more efficient strategy would be to search the set of nonblack things, checking each to see if it is a raven.

More generally, what constitutes an optimal search strategy for falsifying evidence regarding a universal claim depends—other things being equal—on the size of the class defined by the subject of the claim relative to the size of the class defined by the complement of the predicate. In other words, although there is no logical difference between looking for a *P* that is ~*Q* and looking for a ~*Q* that is *P*, psychologically and practically the difference can be substantial. If the set defined by *P* and that defined by ~*Q* are greatly different in size, searching the smaller set will maximise the chances of finding a falsifying case, if one exists, in a given time or for a given expenditure of resources.

If it is the case that for most meaningful conditional assertions that are made for purposes of communication in everyday life the antecedent, *P*, defines a class that is considerably smaller than that defined by the complement of the consequent, ~*Q*, then it is good strategy, as a practical matter, to focus on the former class when seeking to find a member of the intersection of the two classes. This could be interpreted to mean, in effect, making more use of *modus ponens* than of *modus tollens* arguments in attempting to judge the plausibility of conditional claims. This interpretation is similar, at least in spirit, to Cheng and Nisbett's (1993) position that *modus tollens* and the contrapositive are (apparently) not part of the repertoire of people's logical intuitions because they do not have great practical utility.

A COUNTER CASE

Suppose one knew the US senate to be composed, at a particular time (say at the opening of the 104th congress in 1995) of 47 Democrats and 53 Republicans, and of 92 males and 8 females, but that one did not know the parties to which individual senators belonged. Consider the two following claims:

> All Republicans in the US senate are males. (If X is a Republican US senator, X is a male.)
>
> All females in the US senate are Democrats. (If X is a female US senator, X is a Democrat.)

To falsify either of these claims, really the same claim by the rule of the equivalence of contrapositives, one needs to determine that the senate contains at least one female Republican. One could look for such evidence systematically by checking all senators, looking for a female Republican; by checking all Republicans, looking for a female; or by checking all females, looking for a Republican. In view of the relative sizes of the relevant classes, the most efficient approach, in the case of both claims, would be to check all the females looking for a Republican. Given the second claim, this is the strategy we would expect people to select, on the basis of the results that are commonly obtained with Wason's selection task. Given the first claim, selection-task results make us suspect that many people would elect to check all Republicans looking for a female, even though it is not the optimal strategy.

With respect to the first claim, a female Democrat (nonmale non Republican) plays the same role as does the white shoe (nonblack nonraven) in the original statement of Hempel's paradox. However, I suspect that it is much easier to see the observation of a female Democratic senator as confirming of the claim that all Republican senators are male than to see the observation of a white shoe as confirming of the claim that all ravens are black. The difference, I wish to argue, is in the relative sizes of the relevant sets.

When one thinks of universal assertions about the world that one might want to test, my sense is that those that come readily to mind are like "All ravens are black" with respect to the fact that the class defined by the subject tends to be small relative to the class defined by the complement of the predicate. It is possible to find exceptions to this rule, as the example "All Republicans in the US senate are males" attests, but I think these are exceptions and one must work harder to find realistic examples than one does to find examples that fit the rule.

DEGREES OF CONFIRMATION

The idea that a case of a hypothesis supports that hypothesis was identified at the outset as basic to the notion of incremental confirmation. Until now, nothing has been said about the possibility that cases differ with respect to the degree to which they confirm their hypotheses. Intuitively, most of us would probably consider the observation of a case from a small set of possibilities to be more confirming than that of a case from a large set of possibilities. For example, we would be likely to consider the observation of a red marble to be more confirming of the hypothesis that all the marbles in a specified bag are red if the bag contains only three marbles than if it contains 300.

According to Hempel's equivalence condition, an observation that is confirmatory of a given statement is confirmatory also of any logically equivalent statement. On the assumption that confirmatory observations can be confirmatory to different degrees, we might qualify the equivalence condition to the effect that an observation that is confirmatory of a given statement is confirmatory, *to the same degree*, also of any logically equivalent statement.

If one accepts this version of the equivalence condition and the idea that the degree to which cases confirm depends on the sizes of the sets involved, the observation of a white shoe should be taken as confirming of the claim that all nonblack things are nonravens, but only to a very small degree, so it should be considered confirmatory also of the claim that all ravens are black, but only to the same very small degree. According to this line of reasoning, the claim that all ravens are black is confirmed both by the observation of a black raven and by that of a white shoe, as Hempel argues, but the degree of confirmation received is greater—much greater—in the former case than in the latter.

This is one proposed resolution of Hempel's paradox (see e.g. Howson & Urbach, 1989)—not universally accepted. It attributes the *appearance* of paradox to the enormity of the disparity between the size of the class of ravens and that of the class of nonblack objects. According to this view, the observation of a nonblack nonraven (white shoe) should increase the credibility of the belief that all ravens are black but only very very slightly, indeed hardly at all, because the observation of one nonblack nonraven reduces the chance of finding a raven within the set of nonblack things but only by an immeasurably small amount (Hosiasson-Lindembaum, 1940; Mackie, 1963).

AN ILLUSTRATIVE BAYESIAN ANALYSIS

The foregoing ideas can be illustrated with a Bayesian analysis of an imaginary, and very simple, world. Consider a world that contains 1000 things. Suppose it is known that 50 of the things are ravens and that 100 are black. For purposes of this exercise, we should forget what we know about birds in the real world, such as the fact that, as a rule, birds of a given species have the same colour; we should assume that the proportion of ravens in this world that are black could be anything from 0 to 1.0.

Suppose we wish to evaluate the hypothesis that all the ravens are black (H). The Bayesian approach requires that we begin with a set of mutually exclusive and exhaustive hypotheses. In addition to the hypothesis of interest that all ravens are black (H), we must specify at least one contrary hypothesis ($\sim H$) in terms of some proportion, other than 1.0, of ravens that are hypothesised to be black. Application of Bayes's rule will permit us to evaluate the likelihood of one of the hypotheses relative to that of the other.

What to use as $\sim H$ is a question. Should we evaluate the hypothesis that all the ravens in our imaginary world are black against the hypothesis that none is? That half of them are? That all but one of them are? To develop the example, I will use for $\sim H$ the hypothesis that half the ravens are black, but the analysis that follows can be applied to any other value for $\sim H$. The situation for *H: all ravens are black* and *~H: 0.5 of the ravens are black* is shown in Table 4. (For mnemonic convenience, I use R and B to represent ravens and black in what follows, but the reader will note that R and B correspond to P and Q, respectively, of the foregoing.)

Let us assume that, in the absence of any observations, the hypotheses, H and $\sim H$, are equally likely to be true, i.e. initially $p(H) = p(\sim H) = 0.5$. We would like to know how each of the observations that could be made should change our estimates of these probabilities. According to Bayes's theorem, the probability of a hypothesis conditional on an observation, (the posterior probability) is simply the probability of the observation conditional on the hypothesis (conditional probability) multiplied by the prior probability of the hypothesis divided by the sum of such products for all (both) of the hypotheses. Letting D represent an observation, or *datum*,

$$p(H\,|D) = p(D\,|\mathrm{H})p(H)/\{pD\,|\mathrm{H})p(H)+p(D\,|\sim H)p(\sim H)\}. \qquad (1)$$

TABLE 4
Hypotheses H and ~H

	H			*~H*		
	B	*~B*	*T*	*B*	*~B*	*T*
R	50	0	50	25	25	50
~R	50	900	950	75	875	950
T	100	900	1000	100	900	1000

The numbers of combinations of *R* or *~R* with *B* or *~B*, in an imaginary world under the hypothesis (*H*) that all the *R*s are *B* and (*~H*) that half the *R*s are *B*.

The *effect* of a specific observation may be considered the difference between the probability of the hypothesis conditional on the observation and the probability of that hypothesis before the observation was made:

$$\text{Effect of } D \text{ on } H = \mathrm{p}(H\,|D) - \mathrm{p}(H). \tag{2}$$

How we apply the equation for p($H\,|D$) to our imaginary world depends on what we consider an observation, or *datum*, to be. One possibility is to consider the pair of attributes one discovers when one inspects a random object. This gives us four possibilities: *RB*, *R~B*, *~RB*, and *~R~B*. Given the numbers in Table 4, we have:

$$\begin{array}{lcllcl}
\mathrm{p}(RB|H) & = & 0.050 & \mathrm{p}(RB|{\sim}H) & = & 0.025 \\
\mathrm{p}(R{\sim}B|H) & = & 0.000 & \mathrm{p}(R{\sim}B|{\sim}H) & = & 0.025 \\
\mathrm{p}({\sim}RB|H) & = & 0.050 & \mathrm{p}({\sim}RB|{\sim}H) & = & 0.075 \\
\mathrm{p}({\sim}R{\sim}B|H) & = & 0.900 & \mathrm{p}({\sim}R{\sim}B|{\sim}H) & = & 0.875
\end{array}$$

It should be clear that these values are simply the cell entries of Table 4 divided by the total number of objects in the world. The posterior probability of *H* (or of *~H*) given a specified observation is obtained by using these numbers, along with the prior probability of 0.5, in equation (1). So, for example,

$$\begin{aligned}
\mathrm{p}(H{\sim}|RB) &= \mathrm{p}(RB|H)\mathrm{p}(H)/\{\mathrm{p}(RB|H)\mathrm{p}(H)+\mathrm{p}(RB|{\sim}H)\mathrm{p}({\sim}H)\} \\
&= (0.05)(0.5)/\{(0.05)(0.5)+(0.025)(0.5)\} = 0.667
\end{aligned}$$

Computation of the posterior probability of *H* and *~H* for all possible observations yields the following values (the numbers in parentheses represent the effects of the observations, as given by equation 2):

p(*H* |*RB*)	=	0.667	(0.167)	p(~*H* |*RB*)	=	0.333	(–0.167)
p(*H* |*R*~*B*)	=	0.000	(–0.500)	p(~*H* |*R*~*B*)	=	1.000	(0.500)
p(*H* |~*RB*)	=	0.400	(–0.100)	p(~*H* |~*RB*)	=	0.600	(0.100)
p(*H* |~*R*~*B*)	=	0.507	(0.007)	p(~*H* |~*R*~*B*)	=	0.493	(–0.007)

The effects of the various observations on *H* then are as follows. *RB* is the most strongly confirmatory, *R*~*B* is conclusively disconfirmatory, ~*RB* is (contrary to Hempel's claim) somewhat disconfirmatory, and ~*R*~*B* is confirmatory, but only to a very small degree. Inasmuch as *H* and ~*H* are exhaustive and mutually exclusive hypotheses, what is confirmatory (disconfirmatory) for *H* is equally disconfirmatory (confirmatory) for ~*H*.

Of course, if objects are selected for inspection at random, the various possible observations would not be expected to occur with equal frequency; by far the most likely observation would be ~*R*~*B* under either hypothesis, which is why the finding of this pair has relatively little effect on the probability of either hypothesis.

To make the imaginary-world illustration more relevant to Wason's selection task, let us suppose that, with the information in Table 4 in hand, we are told that we can gather information to help us decide whether *H* is true in the following way: we can specify that we would like to inspect a raven (*R*), a nonraven (~*R*), a black thing (*B*), or a nonblack thing (~*B*). If we specify *R*, an *R* will be selected at random from all the *R*s and we will learn whether it is *B* or ~*B*, and similarly for all the other choices.

In what follows, I will use the notation *rB* to represent the case in which one selects a raven and discovers it to be black, and *Rb*, that in which one selects a black thing and finds it to be a raven. In other words, a letter combination represents what one sees, and the lower-case letter identifies the set—selected by the hypothesis evaluator—from which the object was randomly drawn.

Recall that to find p(*RB*|*H*), the number in the cell in Table 4 representing *RB*, under hypothesis *H*, was divided by the total number of objects in the world, 1000. To find p(*rB*|*H*) or p(*r*~*B*|*H*) we divide the number in the cell representing *RB* or that representing *R*~*B*, under hypothesis *H*, by the total number of *R*s in the world; to find p(~*rB*|*H*) or p(~*r*~*B*|*H*) we divide the number in the cell representing ~*RB* or that representing ~*R*~*B*, under hypothesis *H*, by the total number of ~*R*s in the world. Doing this, we obtain the following conditional probabilities:

p(*rB*|*H*)	=	1.000	p(*rB*|~*H*)	=	0.500
p(*r*~*B*|H)	=	0.000	p(*r*~*B*|~*H*)	=	0.500
p(~*rB*|*H*)	=	0.053	p(~*rB*|~*H*)	=	0.079
p(~*r*~*B*|*H*)	=	0.947	p(~*r*~*B*|~*H*)	=	0.921

If we use these numbers in equation 1, again with the prior probability of H set at 0.5,

$$\begin{aligned}\mathrm{p}(H|rB) &= \mathrm{p}(rB|H)\mathrm{p}(H)/\{\mathrm{p}(rB|H)\mathrm{p}(H)+\mathrm{p}(rB|{\sim}H)\mathrm{p}({\sim}H)\}\\ &= (1)(0.5)/\{(1)(0.5)+(0.5)(0.5)\} = 0.667\end{aligned}$$

The same computation applied to all of the preceding conditional probabilities yields:

$\mathrm{p}(H \mid rB)$	=	0.667	(0.167)	$\mathrm{p}({\sim}H \mid rB)$	=	0.333	(–0.167)
$\mathrm{p}(H \mid r{\sim}B)$	=	0.000	(–0.500)	$\mathrm{p}({\sim}H \mid r{\sim}B)$	=	1.000	(0.500)
$\mathrm{p}(H \mid {\sim}rB)$	=	0.400	(–0.100)	$\mathrm{p}({\sim}H \mid {\sim}rB)$	=	0.600	(0.100)
$\mathrm{p}(H \mid {\sim}r{\sim}B)$	=	0.507	(0.007)	$\mathrm{p}({\sim}H \mid {\sim}r{\sim}B)$	=	0.493	(–0.007)

Thus observing a black raven has the same effect on the hypothesis whether the observation results from a random selection from all the objects in the world or from specifically picking a random raven and discovering that it is black. A similar comment applies for all the other possible observations.

Suppose that instead of selecting ravens or nonravens to determine their colour, we selected black or nonblack objects to determine whether or not they are ravens. The conditional probabilities in this case are as follows:

$\mathrm{p}(Rb \mid H)$	=	0.500	$\mathrm{p}(Rb \mid {\sim}H)$	=	0.250
$\mathrm{p}(R{\sim}b \mid H)$	=	0.000	$\mathrm{p}(R{\sim}b \mid {\sim}H)$	=	0.028
$\mathrm{p}({\sim}Rb \mid H)$	=	0.500	$\mathrm{p}({\sim}Rb \mid {\sim}H)$	=	0.750
$\mathrm{p}({\sim}R{\sim}b \mid H)$	=	1.000	$\mathrm{p}({\sim}R{\sim}b \mid {\sim}H)$	=	0.972

The reader may easily verify that use of these numbers in equation 1, again along with $\mathrm{p}(H) = \mathrm{p}({\sim}H) = 0.5$, produces the same posterior probabilities as did use of the preceding set of conditionals.

So, according to this analysis, the effect that a particular observation—say the observation of a black raven—should have on our confidence in the hypothesis that all ravens are black, is the same whether the observation is made as a consequence of selecting an object (at random from among all objects) and discovering it to be a black raven, selecting a raven (at random from among all ravens) and discovering it to be black, or selecting a black object (at random from among all black objects) and discovering it to be a raven. And a similar comment applies to all the other observations that could be made.

In this analysis, I have held the number of black things constant, independently of the truth or falsity of the hypothesis that all ravens are black (Table 4). One might argue that it would be more reasonable to hold the number of black things *other than ravens* constant, thus making the total number of black things (including black ravens) dependent on the status of the hypothesis: in this case there will be more black things in the imagined world if all the ravens in it are black than if only some of them are.

We can give our imaginary world this property by changing the numbers representing the combinations *~RB* and *~R~B* under *~H* from 75 and 875 to 50 and 900, respectively, and modifying the column totals on the *~H* side of the Table to be consistent with these changes. If the foregoing analyses are repeated with the numbers in Table 4 thus modified, it will be seen that no longer is the effect of the observation of a black raven the same whether the observation is made as a consequence of selecting a random raven and discovering it to be black or selecting a random black object and discovering it to be a raven. Selecting a raven and discovering it to be black still changes the probability of the hypothesis that all ravens are black from 0.5 to 0.667, but selecting a black object and discovering it to be a raven changes it only from 0.5 to 0.6. Selecting a nonblack thing and finding it to be a nonraven again increases the probability of *H*, but by a very small amount; in this world, selecting a nonraven and determining its colour has no effect on the probability of *H*, no matter what its colour is found to be.

I will continue to focus on the case in which the total number of black things is held constant. The reader may wish to do the analysis with the number of black nonravens held constant. The main points of this essay are not dependent on which constraint is imposed. And the difference between the implications of the two possibilities is negligible if the number of ravens is assumed to be very small relative to the number of black things in the world.

The question now is, given the task of selecting among *R*, *~R*, *B*, and *~B* to obtain information to help one decide what to believe about the truth or falsity of *H* with respect to our imaginary world, what should we do? Presumably we should like to maximise our chances of observing *R~B* if such an object exists, because this is the only observation that permits a definite conclusion. Beyond that one might argue that our next preference should be for *RB*, because that is next-most influential datum.

Importantly, vis-a-vis the selection task, the probability of making a given observation is highly dependent on the selection choices one makes. The probability of observing a black raven, for example, differs considerably depending on whether one inspects a random raven to see if it is black or inspects a random black thing to see if it is a raven. In our imaginary world, P(*rB*) is 1 under *H* and 0.5 under *~H*, whereas the comparable numbers for p (*Rb*) are 0.5 under *H* and 0.25 under *~H*.

If our objective in the selection task is to select in such a way as to ensure a relatively high likelihood of getting useful information from each inspection, we have to take into account *both* the effect of each of the possible observations and the probability that any given selection will yield that observation. Table 5 summarises what has been said about our imaginary world in this regard so far. The first row of the Table represents what one selects; the second what each

TABLE 5
The Possible Observational Outcomes of Selections

Selection	*Obs*	*Effect of observation on p(H)*	*Prob of observation, given selection*	
			H	*~H*
r	*B*	0.167 (Strongly confirmatory)	1.000	0.500
	~B	–0.500 (Conclusively disconfirmatory)	0.000	0.500
~r	*B*	–0.100 (Moderately disconfirmatory)	0.053	0.079
	~B	0.007 (Weakly confirmatory)	0.947	0.921
b	*R*	0.167 (Strongly confirmatory)	0.500	0.250
	~R	–0.100 (Moderately disconfirmatory)	0.500	0.750
~b	*R*	–0.500 (Conclusively disconfirmatory)	0.000	0.028
	~R	0.007 (Weakly confirmatory)	1.000	0.972

selection could reveal (two possibilities for each selection), the third the effect of what is revealed, and the fourth and fifth the probability of the observation, given the selection, as specified by *H* and *~H*. (The "strongly", "moderately", and "weakly" designations in this Table are intended to connote *relative* strengths only.)

According to most normative theories of choice, a selection should be made on the basis of the expected desirability or "utility" of its outcome, and this is generally taken to be represented by the sum of the utilities of the outcomes the selection could produce, each weighted by the probability of its occurrence. I will not use the concept of expected utility here, because I think the way in which it should be defined in this context is debatable, but will assume that the selection should be based on the expected *impact* of the outcome of that selection on the probability of the hypothesis under consideration, and I will define impact as the *absolute value* of an observation's effect, as represented by equation 2. Table 6 shows the expected impact of the possible selections on the probability of each hypothesis for the world represented by Table 4; the expected impact of a given selection is the sum of the impacts that selection could have, each weighted by the probability of its occurrence.

So, according to this analysis, whether one believes *H* or *~H*, one expects the selection of a raven to be more useful—have greater impact, by a lot, than selection of any of the other possibilities. The ordering of the remaining possibilities depends on whether one believes *H* or *~H*, but, interestingly, selection of a black object (comparable to the even number in Wason's original task) is the second-best choice for both hypotheses.

TABLE 6
Expected Impacts of Selections, Given the Conditions Represented by Table 4

Selection	*Expected impact under hypothesis* *H*	*~H*
r	0.167	0.334
~r	0.012	0.014
b	0.013	0.117
~b	0.001	0.021

Of course the numbers in Tables 5 and 6 follow from the relative sizes of the relevant sets of objects in our imaginary world. These set sizes are intended to represent *ordinally* their relative sizes in the real world, but obviously they do not come close to reflecting the actual relative sizes, at least when objects of interest are ravens and black things. We might move a little in the direction of greater realism by defining a world that contained, say, 1,000,000 objects, 10,000 of which are black and 50 of which are ravens. For this world, again letting *~H* be that half the ravens are black, one obtains the expected impacts of selections in Table 7.

The most striking effect of increasing the differences among the sizes of the various sets in the way indicated is the great reduction of the expected impact of all selections other than *r*. If one accepts this line of reasoning, and assumes that, in most real-world situations of interest, the class represented by *P* is very small relative to that represented by *Q*, and that the latter is very small relative to the universe of discourse, one could convince oneself that there is seldom much point in looking anywhere except among the *P*s for evidence regarding the tenability of the hypothesis that "All *P*s are *Q* or "If *P* then *Q*".

Another way in which these examples could be modified to reflect situations that may be closer to many found in the real world would be to use, as a counter

TABLE 7
Expected Impacts of Selections, Given a World of 1,000,000 Objects, 10,000 of Which are Black and 50 of Which are Ravens

Selection	*Expected impact under hypothesis* *H*	*~H*
r	0.166667	0.333334
~r	0.000012	0.000012
b	0.001432	0.001434
~b	0.000006	0.000019

to the hypothesis that all ravens are black, a hypothesis that, say, 98% of all ravens are black. Or one could start by leaning in one direction or the other, say by giving *H* a high prior probability and *~H* a correspondingly low one. The reader may want to repeat the preceding exercise using sets of numbers that might be considered more realistic than those in this illustration.

I am well aware that the examples considered here do not constitute a proof of anything. I offer them only as evidence of the plausibility of the conjecture that both the strong preference that people show for the *P* alternative in the selection task and their reluctance to accept white shoes as useful evidence of the truth of the claim that "All ravens are black" may have a robust rational basis in the kind of reasoning that practicality demands in the real world.

Legrenzi, Girotto, and Johnson-Laird (1993) have attributed the preference that people typically show for *P* to a tendency to focus on what is explicitly represented in their mental models of the situation; they assume that models of the selection task in its original form usually explicitly represent *P* and, in some cases, *P* and *Q*. The present analysis is not incompatible with this view; one might say it provides a statistical justification for focusing on *P*, whether by means of representation in a mental model or otherwise.

Oaksford and Chater (1994) have proposed a model of the selection task that is similar in principle to this analysis, according to which people should make selections so as to maximise their gain in information regarding the tenability of the hypothesis that the conditional, "If *P* then *Q*", is true relative to the tenability that it is false. They define the information gained from observing the back of a given card in terms of Shannon's measure and argue that, on their assumptions—notably that the properties described by *P* and *Q* are rare—the expected information gain for card selections is ordered as $P > Q > {\sim}Q > {\sim}P$, which corresponds to the ordering that people's selections often follow (and to the ordering produced by the present analysis).

Evans and Over (in press) question the plausibility of Oaksford and Chater's information measure as an indication of the value of what one learns by turning over cards in the selection task, but they support the general idea that selections should be influenced by what one expects to learn from them. For present purposes, more important than a precisely defined measure of information gain (or expected gain) or the exact ordering of cards in terms of their expected informativeness, is the point that it is reasonable to assume that beliefs about the relative commonality of specific entities or properties in the world will translate into different expectancies as to what the turning of specific cards in the selection task will reveal.

The common failure to select both potentially falsifying cards in the selection task has often been interpreted as evidence that many people tend to look only or primarily for directly confirming evidence and not for evidence that would disconfirm or falsify, and that they do not consider looking for falsifying information as an indirect means of confirmation. Persistence in failing to select both potentially falsifying cards even after the falsification principle has been explained (Wason, 1969; Wason & Johnson-Laird, 1972; Wason & Shapiro, 1971), and sometimes even when people have been instructed specifically to attempt to demonstrate the conditional assertion to be false (Wason & Golding, 1974), might seem to support this view.

Because both *P* and ~*Q* have the potential to yield either confirmatory or disconfirmatory information, one must be cautious about taking the selection of either of these as a compelling reason to conclude that the selector is looking for a particular type of evidence. However, the justification for focusing primarily on *P* is very strong. As the preceding analysis shows, if disconfirmatory evidence exists, one is much more likely to find it by inspecting *P*, than by inspecting ~*Q*, on the generally plausible assumption that *P* is much smaller than ~*Q*. And, although either *P* or ~*Q* could conclusively disconfirm the hypothesis, only *P* could yield evidence that is (relatively) strongly confirmatory (*PQ*), whereas ~*Q* could yield confirmatory evidence, but only of the indirect "white-shoe" variety (~*Q*~*P*).

None of this is to deny that when participants in a selection-task experiment interpret their task to be to specify which *of the four cards in view* must be turned over in order to determine the truth or falsity of the assertion *with respect to those four cards*, they should select the card showing *P* and the one showing ~*Q*; this is what logic demands. The evidence that many people do not understand the logic of the situation appears to be quite strong. The point here is that the behaviour that appears to be irrational—perhaps *is* irrational—in the context of the task when interpreted in this way can be seen as being based on principles that make sense when the task is interpreted as that of indicating which of the four *types* of cards represented should be turned over in order to determine the truth or falsity of the assertion *in general*, which seems more representative of the types of hypothesis testing that one is likely to want to do in the everyday world.

SOME EMPIRICAL FINDINGS

The discussion to this point has been largely speculative. The conjecture that the choices people make in the selection task may be determined, in part, by extralogical factors, such as the relative sizes of the sets involved and the presumed relative informativeness of different possible observations in real-world analogs of the laboratory task, is invited by theoretical considerations alone. The same considerations make Hempel's paradox seem less paradoxical than it otherwise might. Thanks to some recent attempts by several investigators to begin to bridge the considerable gap between the literature on reasoning and that on decision making, there are some empirical findings that relate to the conjecture.

The idea that the choices people make should be guided by considerations of subjective expected utility, which is ubiquitous in the literature on decision making, has not been very prominent in the literature on reasoning. However, some investigators of reasoning have begun to argue that performance on reasoning problems, and in particular on the selection task, cannot be accounted for adequately without reference to what are usually thought of as decision-theoretic constructs (Kirby, 1994a,b; Manktelow & Over, 1991, 1992). Oaksford and Chater (1995, p.133) suggest that the selection task “poses a problem of optimal data selection, rather than a problem of logical inference, as is frequently assumed”.

I have already noted that Oaksford and Chater (1994) have proposed a model of the selection task according to which people should make selections so as to maximise their gain in information regarding the tenability of the hypothesis that the conditional under consideration is true relative to the tenability that it is false. Kirby (1994a) makes a distinction between *inferential* and *choice* processes, and contends that both types must be taken into account to explain performance of the selection task. Choices, he suggests, may be influenced by a variety of non-inferential factors such as their likely costs and benefits and certain response biases (matching bias, attentional bias). The fact that one fails to select a logically normative card in any given instance of the selection task is not compelling evidence that one does not understand the *possibility* that that card contains disconfirming information; it could be, he argues, that one simply considers the disconfirming outcome to be unlikely or unimportant.

Among the more important findings, vis-a-vis the speculations in this article, is that of a direct relationship between the size of the set defined by *P* and the probability that participants in the selection task would choose to check the alternative representing *~Q*. Kirby (1994a) hypothesised that—assuming the objective of identifying cards that show the rule under consideration to be false—people may fail to select the *~Q* option because the probability of finding a disconfirming instance by checking the *~Q* alternative may, in many cases, be judged to be small, relative to the probability of finding one by checking the *P* alternative. Reasoning that selection of a specific card should become more likely as the odds of finding disconfirming information on the back of that card increase, he varied the size of the set identified by *P* and found the predicted relationship between the size of this set and the relative frequency with which people selected *~Q*.

Kirby’s interpretation of this result was challenged by Over and Evans (1994) and defended by Kirby (1994b), but the challengers did not argue that outcome probabilities have no effect on selections. They cited results obtained by Pollard and Evans (1983), who made participants aware of the probabilities of combinations on cards through a learning task and then found that *~Q* selections increased with the probability that the conditional assertion was false, as supportive of Kirby’s hypothesis.

Further support for the idea that expectations regarding outcomes can affect selections comes from a study by Love and Kessler (1995) who varied task scenarios and instructions in such a way as to lead participants to have different beliefs about the probability that *P* and ~*Q* would co-occur. The greater the probability of co-occurrence, the more likely participants were to select *P* and ~*Q*. An investigation by Platt and Griggs (1995) of the effect of *P*-set size on probability of selection of ~*Q* yielded negative results; but in this study, Platt and Griggs used explicated statements and had participants give a reason for each of the cards they selected or decided not to select, and these conditions yielded relatively high rates of selection of ~*Q* independently of other factors. The effects of relative set sizes on performance of the selection task seem certain to be the focus of further research.

In the attempt to bring decision-theoretic notions to bear on the study of conditional reasoning, investigators have begun to focus not only on outcome probabilities, but on utilities as well (Kirby, 1994a; Love & Kessler, 1995; Manktelow & Over, 1990, 1991, 1992). The results that have been obtained in this regard are not so relevant to the present discussion, but they demonstrate the importance of utilities to an understanding of conditional reasoning, especially deontic reasoning.

CONFIRMATION OF CONFLICTING CLAIMS BY THE SAME EVIDENCE

As noted earlier, according to the equivalence condition, as stated by Hempel, the observation of a white shoe should be taken as confirmatory evidence both for the claim that all ravens are black and for the contradictory claim that all ravens are red. It may seem obvious that the same data cannot be confirmatory for each of two mutually exclusive hypotheses. Poundstone (1990) takes this position in a discussion of Hempel's paradox. He uses a red herring rather than a white shoe in his argument, but that is an inconsequential difference. He points out that if one accepts Hempel's argument that a red herring should be taken as evidence that all ravens are black, it must also be taken as evidence that all ravens are white, because it confirms the assertion, "All nonwhite things are nonravens", which is the contrapositive of "All ravens are white". "An observation cannot confirm two mutually exclusive hypotheses", Poundstone argues (1990, p.26):

> Once you admit such a patent contradiction, it is possible to "prove" anything. The red herring confirms that the color of all ravens is black, and that that color is white; ergo:
>
> Black is white. QED.

In fact it is not the case that admitting that the same evidence can confirm two mutually exclusive hypotheses necessarily leads to logical chaos. When Bayes's rule is used to evaluate a set of three or more mutually exclusive hypotheses, it is not unusual for more than one of the hypotheses to increase in probability following a given observation. It is easy to see how this can be the case.

Assume that an urn contains red and white balls, that the proportion of red balls is known to be, 0.1, 0.8, or 0.9, and that the three possibilities, which will be denoted H_1, H_2, and H_3, respectively, are assumed to be equally probable, p = 0.333, *a priori*. It will perhaps be intuitively obvious that the random drawing of a red ball should decrease the plausibility of the first hypothesis and increase the plausibility of both of the others; in fact, applying Bayes's rule following the drawing of a single red ball yields posterior probabilities for H_1, H_2, and H_3 of 0.056, 0.444, and 0.500, respectively, which represents an increase for both H_2 and H_3. So at least according to one highly regarded approach to probabilistic reasoning, it is not impossible, or even necessarily unusual, for the same bit of data to be confirmatory of mutually exclusive hypotheses, and our intuitions are not, I think, offended by the example given to illustrate the point.

The idea that observation of a white shoe constitutes evidence that is confirmatory for the hypothesis that all ravens are black may be difficult to accept, for a variety of psychological reasons, but, at least to a Bayesian, the fact that acceptance of this idea requires that one see the same observation as confirmatory with respect to conflicting hypotheses as well should be no impediment. Perhaps even many nonBayesians will find it easy to see, intuitively, how the same data might, in some situations, increase the plausibility of two or more competing hypotheses, only one of which can be correct. To make intuitive sense of this it helps to bear in mind that it is possible to confirm (in the incremental sense, though not in the absolute sense) hypotheses that are not true, and to assume that if an unbiased search for evidence is continued, the confirmation of a false hypothesis is likely to be short-lived.

THE ROLE OF WORLD KNOWLEDGE

The statistical resolution of Hempel's paradox rests on the notion of degrees of confirmation. According to this view, the observation of a white shoe, or any nonblack nonraven, is confirmatory of the claim that all ravens are black, but only to a very small degree relative to the amount of confirmation obtained by the observation of a black raven. Hempel acknowledged that this view may have some merit under certain conditions, but dismissed it as a general resolution, in part because it may not always be easy to determine that the class of ~*Q*s is much more numerous than the class of *P*s.

In fact, however, we often do have knowledge of—or can make plausible assumptions about—the relative sizes of classes of interest. And there can be little doubt that this knowledge, or these assumptions, figure prominently in our reasoning about the world. Moreover, usually the complement of a natural class of interest (class of interest being identified by either the subject or predicate in a universal statement) is not a single natural class but is composed of (is the union of) many natural classes. If this conjecture is true, then when people focus on the natural class that is named by the subject of a universal assertion, rather than on the class defined by the complement of the predicate, they are making efficient use of their cognitive resources.

Our knowledge of the world that is likely to influence our reasoning about classes of objects is not limited to the relative sizes of some of those classes. *A propos* the raven paradox, we know, for example, that birds of a given species usually are coloured alike. Having seen a few birds—or even a single bird—of an unfamiliar species for the first time, we are likely to assume that all birds of that species are coloured as the one(s) we have seen. Thagard and Nisbett (1982/1993) apply this fact to the raven paradox. Starting with the assumption that the degree to which an observation confirms a generalisation in the mind of an observer depends on the observer's background knowledge of variability among the kinds of entities involved, they argue that the observation of a black raven is more confirmatory of the generalisation "All ravens are black" than is the observation of a white shoe, because of the disparity of the observer's knowledge of birds and all things other than black ravens. According to Thagard and Nisbett (1982/1993, p.63):

> Our background knowledge tells us that ravens are kinds of birds, and black is a kind of color, and that birds are fairly invariant with respect to color. However, we have no analogous background knowledge about nonblack things and nonravens. "Nonblack" and "nonraven" are not kinds of anything. With those properties, we are relegated to doing the kind of induction by simple enumeration which requires us to gather very many instances before we can have any confidence that we have more than an accidental correlation. In contrast, "raven" and "black" fit into our knowledge system in such a way that we can use information about variability of kinds to establish a high degree of confirmation on the basis of relatively few instances. This, in addition to size of the relevant classes, allows us to judge that "All ravens are black" is much better confirmed by a black raven than "All nonblack things are nonravens" is confirmed by a white shoe.

CONFIRMATION AND CONDITIONAL REASONING IN EVERYDAY LIFE

In real-world situations, we are less likely, I suspect, to be interested in the truth or falsity of universal statements than in that of near-universal statements. Excepting logical or mathematical tautologies, and fundamental laws of physics, universal statements are not, as a rule, true statements of how things are. It is probably not true, for example, that all ravens are black—that every single raven that exists is black. Almost certainly there exists somewhere an albino raven. More likely there are many of them.

In everyday language, we often use "all" to mean "all or nearly all", or "a very large percentage". And, for practical purposes, the knowledge or assumption that nearly all Xs, or a very large percentage of Xs, are Ys is as useful, or very nearly as useful, as knowing that all of them are.

"All", taken literally, and "nearly all" have quite different implications for confirmation and falsification, however. I would be reluctant to bet on the truth of the claim that literally *all* ravens are black, no matter how many black ravens I had seen without encountering a single one of another colour. My confidence in the truth of the claim that *nearly all*—most, a very high percentage of—ravens are black could get quite high if I had seen a great many ravens all of which were black. Conversely, observation of a single nonblack raven is enough to convince me that the claim that all ravens are black is false; however, the sighting of one or a few nonblack ravens will not necessarily shake my confidence in the truth of the claim that nearly all ravens are black, provided my sample is large enough to be considered representative of the general population and the percentage of the nonblack ravens in it is small.

Also, we cannot, for practical purposes, consider confirmation an all-or-none affair as it applies to everyday matters. Many of the questions that concern us are not questions that we can expect to be able to answer with certainty. The truth or falsity of many of the assertions that we wonder about cannot be determined beyond doubt. Each of us believes some things more strongly than others, and, presumably, rightly so. The evidence of the tenability of beliefs is more compelling in some cases than in others. This is true of beliefs involving universal claims and of those involving near-universal claims as well.

EVERYDAY REASONING, HEMPEL'S PARADOX, AND WASON'S SELECTION TASK

According to Popper's logic of falsifiability, the best way to test the credibility of universal, or near-universal, claims typically is to look for—try to think of—exceptions to them. I believe that we often do just this, and that people have an intuitive grasp of the principle that a single counterexample suffices to show a universal claim to be false. What is more natural than pointing out counterexamples to generalisations we wish to contest? This is not to suggest that everyone who uses the principle is highly conscious of doing so, or can verbalise it in precise terms. However, when we find an exception to a universal claim, we know the claim to be false; in the case of near-universal claims, the easier it is to find exceptions, the less credibility we give to the claims. ("Exception" is a bit of a misnomer as applied to near-universal claims—a white raven would not literally constitute an exception to the claim that nearly all ravens are black—but the term is used loosely here to connote an observation the possibility of which is denied by the universal claim and the probability of which is implied to be low by the near-universal claim.)

Hempel's paradox is seen as a paradox, in part, because looking for an exception to the claim that all ravens are black by searching the class of nonblack things is such an obviously unpromising way of finding a nonblack raven if one exists. Generally, when one has the goal of deciding whether or not to accept a universal, or near-universal, claim, a search is much more likely to turn up exceptions, if they exist, if it is conducted on the class identified by the subject in a categorical statement or on the class identified by the antecedent in a conditional statement; there are exceptions to this rule, but they are exceptions.

Our tendency to focus on these classes is justified, according to this view, by practical considerations. Searching the class identified by the complement of the predicate term of a categorical statement or by the complement of the consequent of a conditional statement would typically be enormously more time consuming and effortful, or, for the same expenditure of time and effort, it would yield a very much smaller return.

But, on the principle that a case of a hypothesis supports that hypothesis, it makes sense to look not only for exceptions to a hypothesised rule, but also for incrementally confirmatory instances. I believe we do that too, and, indeed, that we are more inclined to look for confirmatory cases than for disconfirming ones. It seems safe to assume that for most of the generalisations one would be interested in evaluating in meaningful contexts, the class designated by the subject of a categorical assertion or the antecedent of a conditional is smaller than that designated by the predicate of a categorical assertion or the consequent of a conditional—and therefore more likely to yield a confirmatory case—and very much smaller than the complementary classes, so very much more likely to yield confirmatory evidence than they. This, coupled with the fact that, of the smaller classes, P and Q, only P can yield disconfirming evidence, makes P a more promising class to search, by far, than any of the others.

In sum, when the selection task can reasonably be interpreted as the specification of which of four cards must be turned over in order to determine whether the conditional assertion in question is true in general, the tendency to focus much more on P than on *any* of the other possibilities is quite consistent with—perhaps even predicted by—the assumption that people are searching for information in a relatively efficient, possibly optimal, way. When it is clear that the task is understood to be to indicate which of the cards must be turned over in order to determine the truth or falsity of the assertion in question with respect only to those four cards, the extra-logical considerations do not apply. In this case, the (only) correct response is selection of P and $\sim Q$. When people understand the task in the latter way and, nevertheless, make selections other than P and $\sim Q$, or fail to select the second of these, it could be because they have a poor understanding of logic; however, it could also be that they, somewhat uncritically, carry over to laboratory situations approaches to reasoning problems that have proved to be highly effective in everyday life.

A final conclusion to be drawn from these considerations is that if one wants to get unambiguous results from performance of one or another version of the selection task, it is imperative that it be clear to those who are asked to perform it which of the possible interpretations of the task is intended. In fact there are more interpretations than the two mentioned here (Platt & Griggs, 1995), but distinguishing between the general and specific interpretations suffices for present purposes.

Manuscript received 4 October 1995
Revised manuscript received 31 January 1996

REFERENCES

Cheng, P.W., & Nisbett, R.E. (1993). Pragmatic constraints on causal deduction. In R.E. Nisbett (Ed.), *Rules for reasoning* (pp.207–227). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*. *31*, 187–276.

Evans, J.St.B.T. (1982). *The psychology of deductive reasoning*. London: Routledge & Kegan Paul Ltd.

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.

Evans, J.St.B.T., & Over, D.E. (in press). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*.

Good, I.J. (1967). The white shoe is a red herring. *British Journal of the Philosophy of Science*, *17*, 322. {Reprinted in I.J. Good (1983). *Good thinking: The foundations of probability and its applications* . Minneapolis: University of Minnesota Press.}

Hempel, C.G. (1945). Studies in the logic of confirmation (I). *Mind, 54* (213), 1–26.

Hosiasson-Lindenbaum, J. (1940). On confirmation. *The Journal of Symbolic Logic*, *5*, 133–148.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. La Salle, IL: Open Court

Kirby, K.N. (1994a). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition*, *51*, 1–28.

Kirby, K.N. (1994b). False alarm: A reply to Over and Evans. *Cognition*, *52*, 245–250.

Legrenzi, P., Girotto, V., & Johnson-Laird, P.N. (1993). Focusing in reasoning and decision making. *Cognition*, *49*, 37–66.

Love, R.E., & Kessler, C.M. (1995). Focusing in Wason's selection task: Content and instruction effects. *Thinking and Reasoning*, *1*, 153–182.

Mackie, J.L. (1963). The paradox of confirmation. *British Journal of the Philosophy of Science*, *13*, 265–277.

Manktelow, K.I., & Over, D.E. (1990). Deontic thought and the selection task. In K.J. Gilhooly, M.T.G. Keane, R.H. Logic, & G. Erdos. (Eds.), *Lines of thinking* (Vol. 1). London: Wiley.

Manktelow, K.I., & Over, D.E. (1991). Social roles and utilities in reasoning with deontic conditionals. *Cognition*, *39*, 85–105.

Manktelow, K.L., & Over, D.E. (1992). Utility and deontic reasoning: Some comments on Johnson-Laird and Byrne. *Cognition*, *43*, 183–188.

Nicod, J. (1930). *Foundations of geometry and induction* {Translated by P.P. Wiener}. London: Kegan Paul.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.

Oaksford, M., & Chater, N. (1995). Theories, of reasoning and the computational explanation of everyday inference. *Thinking and Reasoning*, *1*, 121–152.

Over, D.E., & Evans, J.St.B.T. (1994). Hits and misses: Kirby on the selection task. *Cognition*, *52*, 235–243.

Platt, R.D., & Griggs, R.A. (1995). Facilitation and matching bias in the abstract selection task. *Thinking and Reasoning*, *1*, 55–70.

Pollard, P., & Evans, J.St.B.T. (1983). The effect of experimentally contrived experience on reasoning performance. *Psychological Research*, *45*, 287–301.

Popper, K.R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Poundstone, W. (1990). *Labyrinths of reason*. New York: Doubleday.

Rips, L.J., & Marcus, S.L. (1977). Suppositions and the analysis of conditional sentences. In M.A. Just & P.A. Carpenter, (Eds.), *Comprehension processes in comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Rumain, B., Connell, I., & Braine, M.D.S. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults. *If* is not the biconditional. *Developmental Psychology*, *19*, 471–481.

Salmon, W.C. (1973). Confirmation. *Scientific American*, *228*(5), 75–83.

Thagard, P., & Nisbett, R.E. (1993). Variability and confirmation. In R.E. Nisbett (Ed.), *Rules for reasoning* (pp.55–69). Hillsdale, NJ: Lawrence Erlbaum Associates Inc. {Originally published in 1982 in *Philosophical Studies*, *50*, 250–267.}

Tweney, R.D., & Doherty, M.E. (1983). Rationality and the psychology of inference. *Synthese*, *57*, 139–161.

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

Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.

Wason, P.C. (1969). Regression in reasoning? *British Journal of Psychology*, *60*, 471–480.

Wason, P.C., & Golding, E. (1974). The language of inconsistency. *British Journal of Psychology*, *65*, 537–546.

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

Wason P.C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, *23*, 63–71.

Wetherick, N.E. (1993). Human rationality. In K.I. Manktelow & D.E. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp.83–109). London: Routledge.