Psychonomic Bulletin & Review
2003, 10 (2), 289-318

Optimal data selection: Revision, review, and reevaluation

MIKE OAKSFORD
Cardiff University, Cardiff, Wales

and

NICK CHATER
University of Warwick, Coventry, England

Since it first appeared, there has been much research and critical discussion on the theory of optimal data selection as an explanation of Wason's (1966, 1968) selection task (Oaksford & Chater, 1994). In this paper, this literature is reviewed, and the theory of optimal data selection is reevaluated in its light. The information gain model is first located in the current theoretical debate in the psychology of reasoning concerning dual processes in human reasoning. A model comparison exercise is then presented that compares a revised version of the model with its theoretical competitors. Tests of the novel predictions of the model are then reviewed. This section also reviews experiments claimed not to be consistent with optimal data selection. Finally, theoretical criticisms of optimal data selection are discussed. It is argued either that the revised model accounts for them or that they do not stand up under analysis. It is concluded that some version of the optimal data selection model still provides the best account of the selection task. Consequently, the conclusion of Oaksford and Chater's (1994) original rational analysis (Anderson, 1990), that people's hypothesis-testing behavior on this task is rational and well adapted to the environment, still stands.

Since its original appearance in 1994 (Oaksford & Chater, 1994), there has been a great deal of debate and further research on the theory of optimal data selection as an explanation of the behavior observed on versions of Wason's (1966, 1968) selection task. From outside the reasoning domain, the continued fascination with this one task may seem perverse. However, its centrality to the rationality debate—it is the most discussed task in philosophical debates in this area (Cohen, 1981; Stein, 1996; Stich, 1985, 1990)—and the ability to use the paradigm to investigate many different forms of inference (Fiddick, Cosmides, & Tooby, 2000) means that it is likely to continue fascinating reasoning researchers for some years to come. The optimal data selection, or information gain, model explains indicative and causal selection tasks in which participants must select evidence concerning the truth or falsity of a conditional, *if . . . then*, that makes a factual claim about the state of the world—for example, if you turn the key, the car starts. Explaining people's behavior here is of enduring interest because of the generality of the problem: deciding which is the best data to select in order to choose between competing hypotheses. This problem is central to everyday life and to scientific practice.

Correspondence concerning this article should be addressed to M. Oaksford, School of Psychology, Cardiff University, P. O. Box 901, Cardiff CF10 3YG, Wales (e-mail: oaksford@cardiff.ac.uk).

In the indicative form of this task, participants are presented with four cards, each of which has a number on one side and a letter on the other. They can see only one side of each card, and the cards are arranged to reveal an A, a K, a 2, and a 7. The participants are asked to indicate which card or cards they want to turn over to determine whether a rule—for example, *if there is an A on one side then there is a 2 on the other side*—is true or false. According to the normative standard provided by logic, the participants should select the cards that potentially falsify the rule. The rule is false only if an instance can be found that conforms to the antecedent of the rule (A) but not to the consequent (2). By convention, the card showing the true antecedent case (A) is labeled *p*, the false antecedent case (K) *not-p*, the true consequent case (2) *q*, and the false consequent case (7) *not-q*. Thus, for the example rule, a card with an A on one side but without a 2 on the other side is a falsifying *p*, *not-q* instance. Only the A (*p*) and 7 (*not-q*) cards are potentially of this type, and consequently, these are the only cards that participants should ask to be turned over. However, they typically select just the A card or the A and the 2 cards. That is, as compared with the standard provided by formal logic, participants' behavior seems irrational.

The information gain model (Oaksford & Chater, 1994, 1998a) explains this behavior as rational by providing a different normative standard against which to assess participants' performance. This is the theory of optimal data selection in Bayesian statistics (Fedorov,

1972; Lindley, 1956). That such an alternative standard was possible should perhaps have been clear from the ongoing debate in the philosophy of science between Bayesians (Earman, 1992; Horwich, 1982; Howson & Urbach, 1989) and adherents of Popper's (1959) falsificationist methodology (e.g., Miller, 1994). We interpret participants' failure to test hypotheses according to the logic of attempted falsification as suggesting that people are not reasoning logically but are using a Bayesian scheme for hypothesis testing. As we will show in more detail below, according to this Bayesian standard, the most frequent card selections are rational.

As we indicated above, there has been much research and critical discussion of the selection task since the optimal data selection model appeared. There are, moreover, other well-specified models of selection task performance (e.g., Johnson-Laird & Byrne, 1991; Rips, 1994). However, there have been no attempts to directly assess how well each model explains the results on the selection task. The main purpose of this paper is to present the results of a model comparison exercise, where we fitted the various models to the data.

Most of the criticism of the optimal data selection approach has focused on various theoretical arguments (e.g., Evans, 1999; Evans & Over, 1996b; Green, 2000; Laming, 1996; Oberauer, Wilhelm, & Rosas Diaz, 1999). But some empirical results have also been reported that, it is claimed, are inconsistent with the optimal data selection model (e.g., Gebauer & Laming, 1997; Green, 2000; Hardman, 1998; Oberauer et al., 1999). Our second purpose in this paper is to review this literature and to reevaluate the theory of optimal data selection in its light.

Before doing this, however, we will locate the information gain model in the current theoretical debate in the psychology of reasoning—in particular, with respect to the issue of dual processes in reasoning. We then will present a revised model. The reason for doing this is to demonstrate that many of the critical points raised can be readily incorporated into the model and that, when they are, more comprehensive explanations of the data would appear to result.

DUAL PROCESSES IN HUMAN REASONING

Recent research in the psychology of reasoning has converged on the view that there are two processing systems involved in human reasoning. This is a familiar idea (e.g., Evans, 1984), but it has been given new impetus by the finding that some people do reason logically some of the time (Braine & O'Brien, 1998; Stanovich & West, 2000) and that they tend to have higher IQs (Stanovich & West, 2000). This observation has been interpreted as supporting a dual-process view (Evans & Over, 1996a; Stanovich & West, 2000). System 1 processes are automatic, unconscious, and based on implicitly acquired world knowledge. System 2 processes are controlled, analytic, and based on explicitly acquired formal rules. The information gain model is part of a larger program that adopts a probabilistic approach to human reasoning (Chater & Oaksford, 1999b, 2001; Oaksford & Chater, 1998a, 2001). This approach was inspired by the inability of logic-based approaches, both in artificial intelligence and in psychology, to deal with knowledge-rich inferential processes (Chater & Oaksford, 1990; Oaksford & Chater, 1991, 1993, 1995, 1998a). In the probabilistic approach, models such as the information gain model provide computational-level theories of System 1 processes in which the probabilities involved are considered as summary statistics computed over world knowledge (Chater & Oaksford, 2001; Oaksford & Chater, 1998a, 2001). On this view, most reasoning involves only System 1 processes. However, people, especially the more intelligent, may acquire explicit logical rules, either culturally or by explicit tuition. This is consistent with the probabilistic approach, where the possibility has already been raised that some people might use System 2 processes to test conclusions generated by System 1 (Chater & Oaksford, 1999b).

The critical question is the balance of System 1 versus System 2 processes in human reasoning. Most contemporary theorizing is about System 2 processes (e.g., Johnson-Laird & Byrne, 1991; Rips, 1994). However, results from the selection task (Stanovich & West, 1998) suggest that, at most, 10% of university students are capable of engaging System 2 processes when reasoning. If, as this result suggests, most reasoning invokes only System 1 processes, surely this is where reasoning researchers should be looking. A probabilistic approach to these processes explains people's performance in the laboratory as a rational attempt to make sense of the tasks they are set by applying strategies adapted for coping with the uncertainty of the everyday world.

A further goal of this paper, therefore, was to investigate the balance of System 1 versus System 2 processes in human reasoning on the selection task. We achieved this in the model-fitting exercise by comparing System 2 type models, such as mental logics (e.g., Rips, 1994) and mental models (Johnson-Laird & Byrne, 1991), with the System 1 information gain model. We will argue that although these System 2 theories would appear to need to invoke some System 1 processing to explain the data, the converse is not true. That is, the System 1 information gain model can explain the data without needing to invoke System 2 processes. We now present the revised information gain model.

MODEL COMPARISON

In this section, we will compare a revised information gain model with its main theoretical competition from mental models (e.g., Johnson-Laird & Byrne, 1991) and mental logic (e.g., Rips, 1994) theories. We will briefly outline the information gain model, highlighting the revisions made since it first appeared in Oaksford and Chater (1994). We then will outline the mental models and mental logic theories and propose an implementa-

tion in a processing tree model (Batchelder & Riefer, 1999). We then will report the fits of these models to the data on the standard selection task and the negations paradigm selection task (Evans & Lynch, 1973), in which negations are systematically varied in the antecedents and consequents of the task rule.

The Information Gain Model

Oaksford and Chater (1994) characterized a participant's job in the selection task as selecting data to discriminate between two hypotheses. In one of the hypotheses, there is a dependency between the antecedent p and the consequent q of a conditional rule, if p then q. For example, the rule might be, *if you turn the key, the car starts*. This hypothesis, which we call the dependence model (M_D), is represented by the contingency table in the upper half of Table 1. In this table, a is the probability of the antecedent (e.g., the probability that someone turns the key) and b is the probability of the consequent (e.g., the probability that the car starts). This model includes an exceptions parameter, ε, which corresponds to the probability of finding *not-q* given p [$P(\textit{not-}q \mid p)$]—that is, the probability that the car does not start even though the key has been turned. This means that we allow for the possibility that the dependency that participants are testing is not considered to be perfect. This modification was first proposed by Oaksford and Chater (1998b).

The precise form of the dependence model shown in Table 1 was first used to explain biases in conditional inference (Oaksford, Chater, & Larkin, 2000; Hattori, 2002, has also used this model in the selection task). This model is a slight modification of that originally presented in Oaksford and Chater (1994), in that the marginal values remain the same for both hypotheses. In the original model, $P(q)$ varied between models. This modification was motivated by various criticisms made in the literature (Evans & Over, 1996b; Green & Over, 1998) that we will discuss later on.

The other hypothesis, against which M_D is compared, is called the independence model (M_I) where p and q are independent (see the bottom half of Table 1)—that is, turning the key has no effect on the probability of the car's starting. This hypothesis is represented by a contingency table similar to that for M_D, but in which the cell values are simply the products of the corresponding marginal probabilities.

Table 1
Contingency Tables for the Dependence Model (M_D) for a Conditional Rule, *If p Then q*, That May Admit Exceptions (ε) and for the Independence Model (M_I)

	q	*not-q*	
	Dependence Model		
p	$a(1-\varepsilon)$	$a\varepsilon$	a
not-p	$b - a(1-\varepsilon)$	$(1-b) - a\varepsilon$	$1-a$
	b	$1-b$	
	Independence Model		
p	ab	$a(1-b)$	a
not-p	$b(1-a)$	$(1-b)(1-a)$	$1-a$
	b	$1-b$	

Note: $a = P(p)$, $b = P(q)$, and $\varepsilon = P(\textit{not-}q \mid p)$.

What participants want to know is which hypothesis truly describes the disposition of letters and numbers on the cards, and their task is to select the data that will provide the most information about making this discrimination. The most informative data are those that produce the greatest reduction in the uncertainty about which hypothesis is true. This goal first requires calculating how uncertain someone is about which hypothesis is true before he or she selects any data. Intuitively, to quantify uncertainty, a measure is needed that is at a maximum when someone does not know which hypothesis to believe—that is, it is at a maximum when the subjective probability of each hypothesis being true is .5. It should be 0 when he or she is certain that one hypothesis or the other is true. Shannon–Wiener (Shannon & Weaver, 1949; Wiener, 1948) information has exactly these properties, which is why it was used in Oaksford and Chater's (1994) original model:

$$I(\mathrm{M_D}, \mathrm{M_I}) = \sum_i P(\mathrm{M}_i) \log_2 \left(\frac{1}{P(\mathrm{M}_i)} \right). \qquad (1)$$

In Equation 1, $P(\mathrm{M}_i)$ indicates the prior probability that M_D or M_I truly describes the relationship between the letters and numbers on the cards. Someone's uncertainty before selecting any data will be at a maximum of 1 bit when the prior probabilities of the dependence and the independence models are the same—that is, when $P(\mathrm{M_D}) = P(\mathrm{M_I}) = .5$. Moreover, his or her uncertainty will be 0 when either it is known for certain that the dependence model is true [$P(\mathrm{M_I}) = 0$] or it is known for certain that the independence hypothesis is true [$P(\mathrm{M_I}) = 1$]. Thus, Shannon–Wiener information captures our intuitions about a measure of uncertainty.

To determine the amount of information someone gains by turning over a card requires working out how uncertain he or she is after selecting some data. The difference between this uncertainty and the prior uncertainty, calculated in the last paragraph, indicates the gain in information provided by a piece of data (i.e., what is on the other side of a card). Calculating the new uncertainty requires calculating the probability of each model, given some data—that is, $P(\mathrm{M}_i \mid \mathrm{D})$—and these values can be calculated using Bayes' theorem:

$$P(\mathrm{M}_i \mid \mathrm{D}) = \frac{P(\mathrm{D} \mid \mathrm{M}_i) P(\mathrm{M}_i)}{\sum_j P(\mathrm{D} \mid \mathrm{M}_j) P(\mathrm{M}_j)}. \qquad (2)$$

To use Equation 2, the likelihoods of the data given each hypothesis, $P(\mathrm{D} \mid \mathrm{M}_i)$, are needed [we already have $P(\mathrm{M_D}) = P(\mathrm{M_I}) = .5$], and these can all be calculated directly from the contingency tables in Table 1. For example, take the rule, *if there is an A on one side then there is a 2 on the*

other side. Now suppose someone is contemplating turning the A card (p) because he or she thinks there may be a 2 (q) on the back. The probability of finding this piece of data—that is, the 2 on the back of the A card—given the dependence model, $P(q \mid p, \mathrm{M_D})$, is $1 - \varepsilon$. In the independence model, the probability $P(q \mid p, \mathrm{M_I}) = P(q \mid \mathrm{M_I}) = b$. Putting these values into Bayes' theorem (Equation 2) means that the probability that the dependence model is true, given that someone finds a 2 on the back of the A card, $P(\mathrm{M_D} \mid q, p)$, is $(1 - \varepsilon)/(1 - \varepsilon + b)$. And of course, $P(\mathrm{M_I} \mid q, p) = 1 - P(\mathrm{M_D} \mid q, p)$. To determine how uncertain someone is after finding a 2 on the back of the A card requires using these posterior probabilities in Equation 1. We can then calculate the information gain (I_g) associated with turning the A card to find a 2 (p_q):

$$I_g(p_q) = I(\mathrm{M_D}, \mathrm{M_I}) - I(\mathrm{M_D}, \mathrm{M_I} \mid p_q) \tag{3}$$

However, in the selection task, participants never actually get to turn over the cards to see what is on the other side—that is, they never actually get to see the data. Consequently, they must make their decision on whether to turn a card on the information gain they might *expect* to find by turning a card. This requires calculating the posterior information for both possibilities. For example, for the A card (p), the posterior information must be calculated not only for when a 2 (q) is found, but also for when a 7 (*not-q*) is found. The latter is calculated in exactly the same way as we have already outlined. To calculate the gain in information we can expect from turning the A (p) card means that these two posterior information values must be weighted by the probability of finding either a 2 (q) or a 7 (*not-q*). These probabilities are the expected values of either $P(q \mid p)$ or $P(\textit{not-}q \mid p)$ calculated over both models—for example,

$$P(q \mid p) = P(\mathrm{M_D})P(q \mid p, \mathrm{M_D}) + P(\mathrm{M_I})P(q \mid p, \mathrm{M_I}). \tag{4}$$

The expected uncertainty associated with turning the p card [$EI(p)$] is then

$$EI(p) = P(q \mid p)I_g(p_q) + P(-q \mid p)I_g(p_{not-q}), \tag{5}$$

and the expected information gain associated with turning the p card is

$$EI_g(p) = I(\mathrm{M_D}, \mathrm{M_I}) - EI(p). \tag{6}$$

Similar calculations can be performed for the other three cards.

Card choice in the selection task is competitive. So the information gains associated with each card were scaled by the total information available—that is, the information gain summed over the four cards (see Oaksford & Chater, 1998b). So the scaled expected information gain associated with card x is defined as

$$SEI_g(x) = \frac{EI_g(x)}{\sum_{x_i \in \{p, not\text{-}p, q, not\text{-}q\}} EI_g(x_i)}. \tag{7}$$

Hattori (1999) derived a *selection tendency function* (STF) that maps scaled expected information gain onto the predicted probability that a card will be selected. Such a function allows a better comparison of the model's predictions with the actual data. The STF that Hattori chose was a logistic that has also been used to map the outputs of nodes in a neural network onto a probability of responding (e.g., Gluck & Bower, 1988). The probability that any particular card x will be selected to be turned over, $P(\mathrm{T}_x)$, is

$$P(\mathrm{T}_x) = \frac{1}{1 + e^{2.37 - 9.06 SEI_g(x)}}. \tag{8}$$

Hattori (1999) estimated the two parameters in the exponent (2.37 and 9.06) directly from past data on the selection task.

To conclude, the differences between this revised information gain model and the version presented in Oaksford and Chater (1994) are as follows. First, in the dependence model, the probability of the consequent is now the same as in the independence model, as in Oaksford et al. (2000). Second, an exceptions parameter has been introduced, as in Oaksford and Chater (1998b). Third, an STF has been introduced, as in Hattori (1999, 2002).

We show the behavior of the revised model in Figure 1. Each panel represents a card, using a density plot with the probability of the antecedent [$P(p)$] on the x-axis and the probability of the consequent [$P(q)$] on the y-axis. The third dimension, shown by shading, corresponds to the probability that the card should be selected, $P(\mathrm{T}_x)$, according to the information gain model. The lighter the shading, the higher the probability that a card should be selected. As in Oaksford and Chater's (1994) original model, the prior probabilities do not influence the order of the probabilities that each card should be selected, so we set the prior probabilities to be equal—that is, $P(\mathrm{M_D}) = P(\mathrm{M_I}) = .5$. The exceptions parameter, ε, was set to .1. Points in the lower triangular region in black violate the assumption of the dependence model that $P(q) > P(p)(1 - \varepsilon)$.

In modeling performance on the selection task, perhaps the most important point to observe from these density plots is that when the probability of the antecedent, $P(p)$, and that of the consequent, $P(q)$, are both small, there is a region where the probability that the q card should be selected is greater than the probability that the *not-q* card should be selected—that is, $P(\mathrm{T}_q) > P(\mathrm{T}_{not\text{-}q})$. In the selection task, the most frequent response is to select the p and the q cards only. This behavior is usually regarded as irrational. However, according to the information gain model, if people normally regard the probabilities of the antecedent and the consequent as being quite small, this selection of cards is the rational selection: These two cards are more informative about which hypothesis is true, relative to the other cards.

In most tasks that use rules like *if there is an A on one side then there is a 2 on the other side*, it seems clear that the probability of the antecedent [$P(p) = 1/26$] and of the consequent [$P(q) = 1/10$] are both small. However, in most tasks it is never made explicit whether the domain is letter types and single digits or letter tokens (all occurrences of A, B, C, etc.) and all numbers (up to 10, 100?). Consequently, it is rarely clear what these probabilities are.

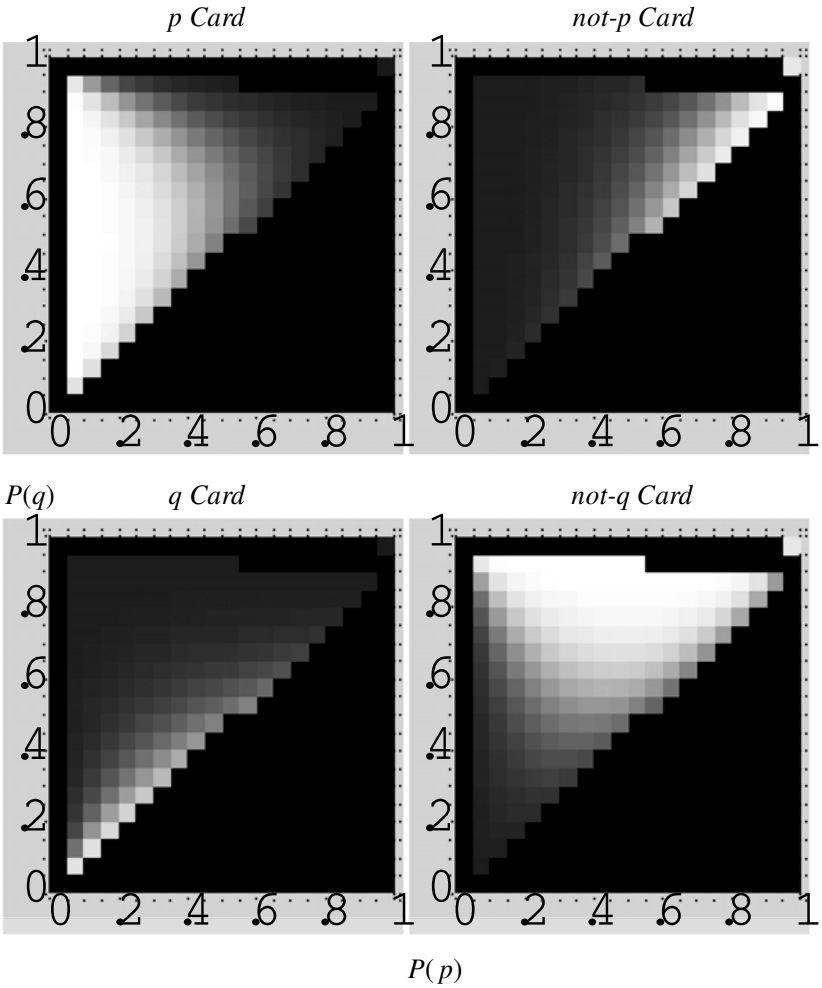

Figure 1. The probabilities with which a card should be selected, $P(T_x)$, as a function of the probabilities of the antecedent [$P(p)$, x-axes] and the consequent [$P(q)$, y-axes] according to the revised information gain model. The lighter the region, the greater the probability that a card should be selected. The prior probabilities [$P(M_I)$ and $P(M_D)$] were set to .5, and the exceptions parameter (ε) was set to .1. The parameters of the selection tendency function were as set in Hattori (1999). Points in the lower triangular region in black violate the assumptions of the dependence model that $P(q) > P(p)(1 - \varepsilon)$.

Under these circumstances, we argue that people revert to prior knowledge. The probabilities of the antecedent and the consequent of a conditional are generally low because the categories of natural language cut the world up quite finely. So, for example, very few things are tables, cars, or gorillas. This is because broad categories—such as, for example, *thing*—that have a high probability of applying to an object in the world are not very useful for telling us what to expect this object to do or how to interact with it. In contrast, knowing that an object is a chair tells us just about everything we need to know. We call the assumption that the categories that function in everyday hypotheses about the world apply only to very small subsets of objects the *rarity assumption.* As we shall see, this assumption seems to explain the experimental results very well.

Mental Logic and Mental Models

The principal theoretical competitors to the probabilistic approach are mental logic (e.g., Rips, 1994) and mental models (e.g., Johnson-Laird & Byrne, 1991) theories. Rips (1994) argued that his PSYCOP model can explain the selection task in the mental logic framework. In PSYCOP, the pattern of logical "errors" is modeled by limiting the number of logical rules and the way that they can be applied. PSYCOP treats each card as an opportunity to use the task rule to draw a conditional inference. There are two logically valid conditional inferences. One is *modus ponens*: *if p then q*, *p*, therefore *q*—for example, all ravens are black, Tweety is a raven, therefore Tweety is black. The other is *modus tollens*: *if p then q*, *not-q*, therefore *not-p*—for example, all ravens

are black, Tweety is not black, therefore Tweety is not a raven. So in the selection task, given the *p* card and the rule *if p then q*, PSYCOP will infer by modus ponens that there should be a *q* on the back of the card. In PSYCOP, modus ponens is implemented by the Forward IF elimination rule. In the selection task, the lack of explicit conclusions (the other sides of the cards) means that PSYCOP cannot apply *backward* rules that work from conclusion to premises (as in a PROLOG interpreter; see Clocksin & Mellish, 1984). These backward rules are required because PSYCOP does not directly implement modus tollens. Consequently, only the *p* card can be selected because this card provides the only match to a rule. According to Rips (1994), some participants also select the *q* card, because they interpret the rule as a biconditional. A biconditional entails its converse—that is, *if p then q and if q then p*. For Rips (1994), succeeding on this problem requires proposing assumptions about what is on the backs of the cards, so that backward rules can also be applied.

Mental models theory (e.g., Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) assumes that people reason logically by manipulating arbitrary mental tokens representing the meaning of the premises. This semantic way of drawing inferences explains selection task behavior largely in terms of people's preference to initially represent only part of the meaning of *if . . . then* statements. This leads people into error if they do not subsequently *flesh out* these representations to express the full meaning of these sentences. So, for example, the rule *if p then q* should be represented by all the instances that make it true. However, people may represent only the named cases—that is,

$$\begin{array}{ll} [p] & q \\ \ldots & \end{array} \qquad (9)$$

where [*p*] means that *p* is exhausted and, so, any other instances must be associated with *not-p* and ". . ." is an ellipsis indicating that other unrepresented models may be relevant. Each line in this representation corresponds to a line in the truth table representing the conditional. Equation 9 shows just the case in which the antecedent *p* is true and *q* is true. Because only the *p* has a value on the other side bearing on the truth or falsity of the rule, people turn this card but not the *q* card. However, if participants believe the rule to be a biconditional, where both *if p then q* and *if q then p* are true, represented as

$$\begin{array}{ll} [p] & [q] \\ \ldots & \end{array} \qquad (10)$$

they will turn both cards. Moreover, if they flesh out their model to include other cases that make the rule true—that is,

$$\begin{array}{ll} [p] & q \\ not\text{-}p & q \\ not\text{-}p & not\text{-}q \end{array} \qquad (11)$$

—people will realize they must turn the *not-q* card as well, for if the rule is true this card *must* have a *not-p* on the other side.

Both theories allow only four possible response patterns: If 1 is used to indicate *turn* and 0 to indicate *do not turn*, then in the card order *p*, *not-p*, *q*, *not-q*, these are 1000, 1001, 1010, and 1111. If participants adopt a conditional interpretation, modus ponens or the mental model in Equation 9 licenses turning just the *p* card—that is, 1000. However, if they consider what is on the other side or flesh out, as in Equation 11, they should turn both the *p* card and the *not-q* card—that is, 1001. If they adopt a biconditional interpretation, modus ponens, or the mental model in Equation 10 indicates that they should turn the *p* card and the *q* card—that is, 1010. However, if they consider what is on the other side or flesh out, they should turn all four cards—that is, 1111. Thus, both theories make identical predictions. They simply disagree on the precise processes by which people arrive at these different interpretations. However, they both involve two choice points in processing this information. The first is whether a conditional or a biconditional interpretation is adopted; the second is whether the representation is fleshed out (mental models) or participants consider what is on the other side of a card (mental logic). Consequently, we can capture both accounts in a simple processing tree model, shown in Figure 2 (Batchelder & Riefer, 1999).

This model parameterizes these accounts in terms of two probabilities: the probability that people adopt a conditional interpretation (P_C) and the probability that people flesh out their initial representation or consider the other sides of the cards (P_F). In such a model, P_C and P_F are assumed to be independent. This is certainly the simplest instantiation of the model. Moreover, no one, to our knowledge, has ever proposed that these are not independent, although investigating models where the independence assumption is not made might be of interest.

The probability that people adopt the biconditional interpretation is, then, $1 - P_C$, and the probability that they do not flesh out or consider the other sides of the cards is $1 - P_F$. Expressions can then be easily derived for the probability that a particular card will be chosen. First, the *p* card will be chosen regardless of interpretation or whether people flesh out their initial representation or consider the other sides of the cards. So the probability of selecting the *p* card, $P(p\text{-card})$, is

$$P(p\text{-card}) = 1. \qquad (12)$$

Although the mental models theory predicts that the *p* card should always be chosen, we allow for the possibility of error in fitting the model to the data (see below). Second, the *not-p* card will be chosen only if people adopt the biconditional interpretation and flesh out their initial representation or consider the other sides of the cards. So the probability of selecting the *not-p* card, $P(not\text{-}p\text{-card})$, is

$$P(not\text{-}p\text{-card}) = P_F(1 - P_C). \qquad (13)$$

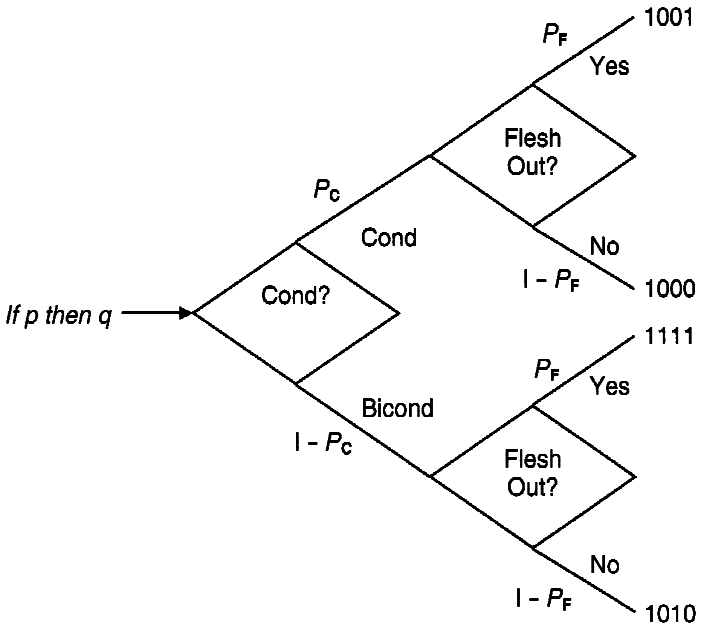

Figure 2. Processing tree model for the mental logic/models theory of the selection task. Cond, conditional; Bicond, biconditional.

Third, the *q* card will be chosen when people adopt the biconditional interpretation, regardless of whether they flesh out their initial representation or consider the other sides of the cards. So the probability of selecting the *q* card, $P(q\text{-card})$, is

$$P(q\text{-card}) = 1 - P_C. \tag{14}$$

Finally, the *not-q* card will be chosen when people flesh out their initial representation or consider the other sides of the cards, regardless of which interpretation they initially adopt. So the probability of selecting the *not-q* card, $P(not\text{-}q\text{-card})$, is

$$P(not\text{-}q\text{-card}) = P_F. \tag{15}$$

This model assumes that the probability of fleshing out a conditional is the same as the probability of fleshing out a biconditional. This is an assumption that could be relaxed, but only at the expense of introducing a further free parameter. In fitting the information gain model to the data, only $P(p)$ and $P(q)$ were free to vary. In model fitting, the greater the number of free parameters, the greater the likelihood of obtaining a good fit. Consequently, to keep the model comparison exercise fair and to avoid making corrections for one model's having more free parameters than another (see, e.g., Sakamoto & Aikake, 1978), it was decided to use model versions that allowed the same number of free parameters.

Model Fitting I: The Standard Abstract Results

In this section, we will report the results of comparing the performance of the revised information gain model with that of the mental models and mental logic theories. We first compared the fit of these models to the data that were used in Oaksford and Chater's (1994) meta-analysis of the abstract data. This meta-analysis (Wolf, 1986) included 13 studies reporting 34 standard abstract selection tasks, involving 845 participants. This model-fitting exercise also addresses the objection that optimal data selection accounts have provided only ordinal fits to the data (Laming, 1996). Hattori (2002) has provided similar fits for a subset of these results. However, the main purpose of this section is to compare the information gain model with alternative theories of the selection task.

In fitting these models to the data, we looked for the values of $P(p)$ and $P(q)$ or P_C or P_F that maximized the log of the likelihood (L) of the data, given either model. Hattori (1999, 2002) adopted the same approach. Because the data will have a joint binomial distribution, L is

$$L = \prod_{j=1}^{J} \binom{F_j}{f_j} p_j^{f_j} (1 - p_j)^{F_j - f_j}, \tag{16}$$

where J is the number of cards (i.e., 4), f_j is the frequency of card selections, F_j is the total number of responses (i.e., N), and p_j is the probability of selecting a card according to the models we are comparing. To estimate best-fitting parameter values, we maximized the log of Equation 16, using a steepest descent search implemented in Mathematica 4.0 (Wolfram, 1999). This was combined with a grid search procedure to ensure a global maximum (Loehle, 2000). The log-likelihood ratio test statistic G^2, which is asymptotically distributed as χ^2, was used to assess the goodness of fit (Read & Cressie, 1988). This statistic evaluates the model fit by comparing the predicted values with a saturated model in which all the val-

ues of p_j are set to the empirically observed proportions of cards selected. As within each experiment, there are four data points, but only two parameters were estimated from the data; G^2 was assessed against two degrees of freedom. For model comparisons, the conventional 5% level of significance is regarded as unreasonably large (Read & Cressie, 1988). The level of significance for rejection was therefore set at the 1% level.

Before discussing the model fits, there are three points we need to make. First, in fitting these models, if one or more of the best-fitting parameter values is less than 0 or greater than 1, we may also reject the model for these data, since this makes no psychological sense. To avoid this happening, the models were always fitted to the data with the constraint that the parameter values must fall in the 0–1 probability interval. Second, we are fitting these models to pooled data, which might be misleading. However, given the nature of the data, there was little choice in this. Third, even though both models have the same number of free parameters, one could be more flexible than the other (Myung & Pitt, 1997)—that is, it can fit almost any data of this type. Since we are proposing that the information gain model is to be preferred, the onus is on us to show that the information gain model does not win just by being more flexible. As an illustration, we looked at the situation in which all cards are selected. This situation is trivial for the mental models theory to model: A perfect fit is obtained when $P_C = 0$ and $P_F = 1$. However, the best fit that the information gain model can achieve is when $P(p) = P(q) = .5$ ($M_D = .5$ and $\varepsilon = .1$, as in all the model fits below). When this is the case, $G^2(2) = 135.65, p < .0001$. That is, the model can be unequivocally rejected. So there are possible results that, although consistent with the mental models theory, cannot be modeled by the information gain theory. Consequently, should the information gain theory provide better fits, this is not simply because it is more flexible, but because people do not tend to behave in ways that are inconsistent with it.

Information gain. In fitting the revised model to these data, we always kept prior probabilities the same at .5—that is, $P(M_D) = P(M_I) = .5$ and $\varepsilon = .1$. Similarly, the parameters of the selection tendency function were kept at the values estimated by Hattori (1999). Indeed, these values were kept fixed in modeling all the results we report in this paper. The results are shown in Table A1 in the Appendix. For 33 out of the 34 experiments reported by Oaksford and Chater (1994), the model could not be rejected—that is, the saturated model did not provide a significantly better fit. For these 34 studies, the average $G^2(2) = 2.60$ ($SD = 2.55$), $p > .20$. Moreover, overall—that is, across all 34 experiments—the model could not be rejected even if the conventional, but unreasonably large (Read & Cressie, 1988), 5% level of significance is adhered to for the goodness-of-fit test [$G^2(68) = 88.37$].

Across all 34 studies, the mean value of the probability of the p card, $P(p)$, was .22 ($SE = .019$), and the mean value of the probability of the q card, $P(q)$, was .27 ($SE = .022$). $P(p)$ was significantly lower than $P(q)$ [$t(32) = 6.61, p < .0001$]. It is worth noting that these values are very close to the expected prior probability of a cause (.25) and of an effect (.27) found by Anderson (1990) when modeling causal estimation tasks (Schustack & Sternberg, 1981). It would seem that participants bring very similar prior expectations to bear in both tasks. This suggests that in the selection task, people by default adopt values for the probability of the antecedent and consequent that are analogous to their knowledge of causes and effects. This is unsurprising when it is considered that, in linguistics, the conditional construction is often assumed to have been introduced into human languages to describe causal structure (Comrie, 1986). In sum, the model can provide quite accurate fits to the data.

Mental logic and mental models. We first note that on any interpretation, the p card should be chosen. Therefore, the probability of selecting this card should be 1. We assessed whether this was the case in the data reported in Oaksford and Chater's (1994) meta-analysis. For each study, we performed a binomial test with p_j set to .99, which provided the probability of the data, assuming that the probability of selecting this card was close to 1 (Siegel & Castellan, 1988). We then combined these probabilities across studies using Fisher's combined test (Wolf, 1986), which yields a χ^2 statistic with $2N$ degrees of freedom, where N = the number of studies. Across the 34 studies, the data differed significantly from what would be expected assuming that the probability of selecting the p card is close to 1 [$\chi^2(68) = 422.52, p < .0001$]. Therefore, this prediction of the mental logic and mental models theories fails.

We then fitted the model to the data. One might argue that the proportion of participants adopting each interpretation is available in the data we are modeling. Consequently, there is no need to estimate these proportions from the card selection frequencies. However, the analysis for the p card that we just reported indicates that people adopt more than the four interpretations countenanced by the mental logic and mental models theories. These theories must regard these interpretations as errors. By fitting the model to the data, we obtain an index of how important these errors are. If good fits can be obtained while maintaining just these four interpretations, the errors are not significant.

However, given the fact that the p card is less than universally accepted, contrary to the prediction of the mental models and mental logic theories, good fits to the data seem unlikely. Indeed, if these theories are fitted to the data with $P(p\text{-card}) = .99$, they can be rejected for 33 of the 34 studies in Oaksford and Chater's (1994) meta-analysis. We therefore sought a fairer comparison. One possibility was to allow the probability of the p card to vary. However, varying the probability of the p card would involve introducing another free parameter, which would violate the constraint that both models should have the same number of free parameters. Another possibility was to allow $P(p\text{-card})$ to be semi-free—that is,

it could be fixed for all 34 data sets at the value that provides the optimal fit. There are two justifications for doing this. First, the information gain model has parameters that have been set globally, although $P(M_D)$ and ε were not optimized in any way. Second, it may be unreasonable to expect people to display so little error (1% in the above analysis) in their selections of the *p* card. Because this parameter applies only to *p* card selections, the maximum likelihood estimate is given by the mean value of the proportion of these cards selected over the 34 studies in the meta-analysis. We therefore computed the fits with the probability of the *p* card, $P(p\text{-card})$, set to this value—that is, .8884. It is these fits that are shown in the Appendix, Table A1. As for the information gain model, G^2 was evaluated against two degrees of freedom.

As for the information gain model, across all 34 studies, the mental models and mental logic theories could be rejected for only one experiment. The average $G^2(2)$ across all studies was 3.26 (SD = 2.79), $p > .01$. However, across all 34 experiments, the mental model and mental logic theories could be rejected [$G^2(68) = 110.99$, $p < .001$]—that is, overall, the saturated model still provided a significantly better fit. In sum, it would appear that, even after relaxing some strong assumptions, the mental models and mental logic theories cannot provide as good fits to the data as the information gain theory.

Discussion. This attempt to fit the mental logic and mental models theories to the selection task data also raises some important theoretical issues. There is no theoretical justification within either mental models theory or mental logic theory for allowing the level of error that we have had to introduce to account for the less than universal selection of the *p* card. According to both theories, the *p* card should always be selected. There are three possible explanations for this level of error. First, the error arises because of response biases: Some participants, some of the time, do not attend to the task demands. Consequently, these participants may simply guess (see Rips, 1994, for a similar explanation of errors in syllogistic reasoning). Second, mental models or mental logic representations must be implemented in neurones that are inherently noisy, and this may lead all participants to make some errors some of the time. Third, a perhaps more theoretically motivated explanation is that the mental representations underlying deductive inference in both mental logic and mental models depend on prior knowledge (Oaksford, 1989; Oaksford & Chater, 1991, 1993, 1995; Schroyens, Schaeken, Fias, & d'Ydewalle, 2000). That is, they rely on System 1 processes. This seems more consistent with the fact that the effects for the *p* card were systematic—that is, selections were significantly less than 100%, rather than random. We concentrate on recent proposals within the mental models framework to illustrate this possibility, but similar considerations may apply to mental logic.

One reason why the *p* card may not be universally selected is that everyday conditionals are affected by prior knowledge. For example, take the rule *if I turn the key, the car starts.* Someone is less likely to infer that the car starts, knowing that the key has been turned, if they also suspect that there is no fuel in the tank. Such knowledge is known to reduce inferences by modus ponens and by modus tollens (e.g., Byrne, 1989; Cummins, Lubart, Alksnis, & Rist, 1991). Moreover, Schroyens et al. (2000) have recently argued that even abstract materials seem prone to these effects. And Feeney and Handley (2000) have introduced similar manipulations into the selection task and have shown marked effects on card selection. In mental models theory, this additional information is represented by a conjunctive antecedent—that is, the further condition that the fuel tank must not be empty is also represented in the mental model. If we assume that even abstract materials must allow such exceptions to a rule, this may explain the nonuniversal acceptance of the *p* card.

However, there is a problem. According to this account there should be similar changes in *not-q* card selections. But there is no correlation in the 34 studies analyzed above between *p* card and *not-q* card selections [$r(32) = .16$, $p = .38$]. Moreover, this fact points to a theoretical problem that also applies to the first two accounts of why the *p* card is not universally selected: random error and noisy neural representations. All these accounts must predict similar effects on the other three cards. However, at the moment, the best fit we can obtain is one in which the frequencies of *not-p*, *q*, and *not-q* card selections are almost perfectly explained by different proportions of participants' adopting the four interpretations allowed by the mental logic and mental models theories. Only the *p* card is affected by these other factors. This is theoretically incoherent. At least prima facie, it seems that the factors we have discussed must influence all card selections, not just the *p* card. Moreover, the work of Schroyens et al. (2000) and Feeney and Handley (2000) seem to indicate that these factors are systematic and nonrandom and can be brought under experimental control. Consequently, the first two options can be discounted: The effects for the *p* card are best regarded as revealing the systematic effects of prior world knowledge—that is, of System 1 processes—on people's data selection behavior.

A question that naturally arises is, does such an account need to be mediated by System 2—that is, analytic or rule-governed—processes of the type suggested by mental models and mental logic theories? We have consistently argued that the probabilities in our probabilistic approach are derived from prior world knowledge, perhaps implemented as activation levels in a neural network (Chater & Oaksford, 2001; Oaksford & Chater, 1993, 1995, 1998a, 2001). Probabilistic accounts of the conditional (Adams, 1966, 1975; Pearl, 1988, 2000) avoid many of the problems of accounting for world knowledge effects that confront extensions of standard logic (see Chater & Oaksford, 2001; Oaksford & Chater, 1998a, 2001). Consequently, our probabilistic approach in the information gain model can be seen as an attempt

to explain selection task performance by using only the effects of prior knowledge unmediated by logical or logic-like representations of the type postulated by mental logic and mental models theories. Or in other words, the information gain theory is a System 1 theory that suggests that little or no System 2 processes are required to explain the data. The model-fitting exercise reported here seems to show that such an account can explain the data better than the latter approaches. If mental logic and mental models theories were supplemented with mechanisms for explaining how all card selections are affected by probabilistic world knowledge, the number of parameters involved can only increase. This would improve the fit, but only at the cost of proposing a theory that is more complicated than the data it is trying to explain. Theoretical parsimony argues against going down this route.

Summary. In sum, the revised information gain model provides very good fits to the data included in Oaksford and Chater's (1994) meta-analysis. This theory also compares favorably with the mental logic and mental models theories. These approaches could not provide as good a fit as the information gain model until we allowed the probability that the *p* card is selected to take on its optimal value. Even then, overall, the mental logic and mental models approaches could be rejected. Finding a principled reason for the nonuniversal selection of the *p* card led us to consider the effects of prior knowledge. It was clear that such effects should occur for all cards. If this is allowed, the question arises as to whether mental logic or mental models representations are really required to explain these data. The good fits for the information gain model, where it is assumed that the relevant probabilities are derived directly from prior world knowledge, seems to suggest that this level of mental representation is not required to explain the selection task.

It could be argued that the information gain theory does better than mental logic or mental models because we are fitting the model to aggregate data. Given the categorical nature of the data, there is little choice about this. However, it could be argued that the mental models and mental logic accounts at least give an idea of what processes are going on in the minds of each individual whose data makes up the aggregate in a way that the information gain account does not. That is, participants adopt different interpretations that are consistent with different response patterns.

This point goes to the heart of the frequently repeated criticism of optimal data selection models, that they do not provide a truly psychological account of the selection task (e.g., Evans & Over, 1996a). That is, they do not provide an account of the mental processes underlying people's performance. However, they do provide an account of what those processes must compute. And as we have argued (see also Anderson, 1990; Marr, 1982), this is a crucial first step in specifying the algorithms that implement human reasoning in the mind. We have also argued that the information gain theory could be implemented in essentially stochastic algorithms such as neural networks (Chater, 1995; McClelland, 1998) or Bayesian networks (Pearl, 1988, 2000). According to such an account, probabilities of responding are related to the activation levels of nodes representing information relevant to card selection. Whether a card is selected requires a statistical decision with inherently noisy neural representations. This account makes a rather direct prediction that contrasts with mental logic and mental models theories. If participants were given many data points to assess, we would expect results within individuals similar to those we see in the aggregate data when participants see only one of each type of data. That is, for a given rule, we would expect individual participants to select a card on some trials, but not on others. The only way the mental logic or mental models theories can make this prediction is if they assume that people change their interpretation of the task rule from trial to trial. This seems rather unlikely. Later on, we will consider some data (Green & Over, 1997; Oaksford & Wakefield, 2003) with which this hypothesis can be tested.

Model Fitting II: The Negations Paradigm

We now turn to fitting these models to Evans's negations paradigm in the selection task (Evans & Lynch, 1973), where negations (*not*) are systematically varied in the antecedent and consequent of a conditional rule. This creates four further rules—*if p then q*, *if p then not-q*, *if not-p then q*, and *if not-p then not-q*—that produce changes in card selections more consistent with falsification (Evans, 1983, 1984, 1989; Evans & Lynch, 1973). This happens for rules with a negated consequent: *if p then not-q*, and *if not-p then not-q*. For these rules, participants select more consequent cards that can make the rule false (i.e., false consequent, or FC, cards) than consequent cards that can make it true (i.e., true consequent, or TC, cards). For example, for the *if p then not-q* rule, participants select the *q* card (FC). Evans (e.g., 1989) explains this finding by people *matching*. That is, participants ignore the negations and match the items named in the rule to the corresponding cards. Thus, it requires no sudden insight into logic to explain why, for example, people select the *q* card for the *if p then not-q* rule. Because the cards that falsify or confirm vary between rules, the convention has been adopted of referring to the cards in the negations paradigm by using the labels true antecedent (TA), false antecedent (FA), TC, and FC. For example, for the *if not-p then not-q* rule, TA is the *not-p* card, FA is the *p* card, TC is the *not-q* card, and FC is the *q* card.

According to Oaksford and Chater (1994), the effects in the negations paradigm can be rationally explained on the assumption that negations define higher probability contrast sets (Oaksford & Stenning, 1992). So for example, the probability that you are not drinking a glass of whiskey is far higher than the probability that you are. This leads to the prediction that if the information gain model is fitted to each condition in the negations paradigm, the values of $P(\text{TA})$ and $P(\text{TC})$ should be higher

when they correspond to negated categories. Before comparing models, we assessed this prediction by fitting the model for each rule individually.

Information gain: Individual rule fits. We fitted the model to the six negations paradigm experiments analyzed in Oaksford and Chater (1994). There were 4 rule conditions in each experiment, making 24 conditions modeled overall. We fitted the model to the data in exactly the same way as for the standard abstract results. The results of the model-fitting exercise are shown in Table A2 in the Appendix. The model could not be rejected for any of the 24 conditions [mean $G^2(2) = 2.61$, $SD = 2.06$, $p > .20$]. Moreover, across all conditions, the model could not be rejected [$G^2(48) = 62.69$, $p > .05$]. We also looked at the fit by rule type. The model could not be rejected for any rule type at the .01 level: For the *if p then q* rule, $G^2(12) = 9.75$, $p > .20$; for the *if p then not-q* rule, $G^2(12) = 23.16$, $p > .01$; for the *if not-p then q* rule, $G^2(12) = 16.07$, $p > .10$; and for the *if not-p then not-q* rule, $G^2(12) = 13.68$, $p > .20$. In sum, as for the standard data, the model provided good fits to the negations paradigm data.

We also checked whether the best-fitting parameter values behaved as predicted by the contrast set account of negation. The means of the best-fitting values for the probability of the antecedent, P(TA), and the probability of the consequent, P(TC), for each rule type were the following: for the *if p then q* rule, P(TA) = .30 (SD = .04) and P(TC) = .34 (SD = .03); for the *if p then not-q* rule, P(TA) = .28 (SD = .08) and P(TC) = .60 (SD = .08); for the *if not-p then q* rule, P(TA) = .37 (SD = .05) and P(TC) = .38 (SD = .07); and for the *if not-p then not-q* rule, P(TA) = .38 (SD = .09) and P(TC) = .46 (SD = .09). When the antecedent was negative, P(TA) was significantly higher than when it was affirmative [$t(22) = 3.29$, $p < .005$; negative, mean P(TA) = .38, SD = .07; affirmative, mean P(TA) = .29, SD = .06]. When the consequent was negative, P(TC) was significantly higher than when it was affirmative [$t(22) = 4.81$, $p < .0001$; negative, mean P(TC) = .53, SD = .11; affirmative, mean P(TC) = .36, SD = .05]. That is, consistent with the contrast set account of negation (Oaksford & Chater, 1994; Oaksford & Stenning, 1992), negated constituents corresponded to higher probability categories.

The *if not-p then q* rule is anomalous. The reason is that, if the contrast set account of negation is correct, the set of things that are *not-p* (e.g., nonwhite cars) is far larger than the set of things that are *q* (e.g., Fords). Consequently, a hypothesis like *all nonwhite cars are Fords* is known at the outset to be false, because of people's experience of nonwhite Nissans, BMWs, Peugeots, and so on (Oaksford, 1998; Oaksford & Chater, 1994). In the model fits, the best-fitting parameter values for this rule were similar to the *if p then q* rule. We adopt the convention that rules are described using ordered pairs $<P(p), P(q)>$ so that, for example, LH means a low-probability antecedent and a high-probability consequent rule. So, in the selection task, the HL rule appears to be treated like an LL rule. We have argued that people tend to revise the probability of the antecedent, $P(p)$, down for the HL rule (Oaksford & Chater, 1994, 1998b). Later on, we will discuss some empirical evidence (Oaksford & Wakefield, 2003) that suggests that we were mistaken and that people seem to compensate for this rule by revising up the exceptions parameter ε. However, for the purpose of fitting the model to the data, we kept ε fixed at .1, as for the other rules, because not doing so would amount to introducing a further free parameter.

Fitting the model to each negations paradigm experiment on an individual rule basis is the same as using eight parameters to model 16 data points—that is, a P(TA) and P(TC) parameter for each of the four rule types in the negations paradigm. Consequently, we will now report model fits for which we fitted the models to all 16 data points in each experiment by adding only a single further parameter.

Information gain: Overall fits. Modifying the information gain model was straightforward. We simply assumed that when the antecedent or the consequent of a conditional is negated—that is, in the *if not-p then q*, *if p then not-q*, and *if not-p then not-q* rules—$P(p)$, $P(q)$, or both are raised by some fixed probability, $P(n)$. So when negated, P(TA) = $P(p) + P(n)$ and P(TC) = $P(q) + P(n)$. As for the standard abstract data, we set the prior probability that the independence model is true [$P(M_I)$] to .5 and the exceptions parameter, ε, to .1. Because there were 16 data points and only three free parameters in each model, G^2 was assessed against 13 degrees of freedom. The results of the model fits are shown in Table A3 in the Appendix. The model could be rejected only for one of the six studies: Manktelow and Evans (1979, Experiment 2). The average $G^2(13) = 22.67$, $p > .02$. However, overall the model could be rejected [$G^2(78) = 136.04$, $p < .01$].

Mental logic and mental models. The mental logic approach has not addressed results from the negations paradigm, so in assessing the fit of these approaches we will concentrate on proposals from mental models theory only. However, as Evans and Handley (1999) have argued, it would seem that similar proposals could be implemented in a mental logic approach.

There have been a variety of proposals to account for the effects of negation on the selection task. First, Evans and Lynch (1973) proposed that people simply match the items named in the rules, rather than doing any reasoning at all. So given a rule *if A then not 2*, participants still select the A and the 2 cards. This was justified from pragmatic considerations concerning the topic of a negated sentence—for example, the topic of *the train was not late* is still the lateness of trains. Note that according to such an account, there is a processing benefit for cards that match items mentioned in the rule. Such an account could not explain all the variation in performance, however, because *p* card selections still dominate and people do select cards other than those that match.[1] Evans (1983, 1984) therefore proposed that pragmatic

heuristics supplement an analytic reasoning component, perhaps provided by mental models. Moreover, Johnson-Laird and Byrne (1991) suggested that this heuristic may be implemented in mental models. When representing a rule such as *if A then not 2*, they suggested that people supplement it with a representation of the matching case:

$$\begin{array}{ll} [\mathrm{A}] & \text{not-2} \\ & 2 \\ & \cdots \end{array} \qquad (17)$$

Again, this means that there is a processing benefit for cards that match named items.

As we have discussed above, Oaksford and Stenning (1992) argued that the difficulty lies with constructing contrast classes for the negated constituents in the task rule. So there is a processing cost for recognizing that nonmatching cards need to be selected. More recently, Evans, Clibbens, and Rood (1996) proposed that the problem relates to the use of implicit negations on the cards—for example, "7" is used to indicate a card that is "not-2." Again, this suggests that there is a processing cost for recognizing that nonmatching cards need to be selected. The latter approaches are difficult to distinguish. Experimentally, Evans et al. (1996) have shown that using explicit negations on the cards (i.e., "not-2" rather than "7") removes matching. Moreover, Oaksford and Stenning have shown that making contrast class construction easier also removes matching. Evans prefers to see his implicit negations account as supplementing mental models, whereas Oaksford and Chater (1994) used the contrast class approach to supplement their probabilistic theory. However, either approach could have been used to supplement mental models (see, e.g., Schroyens et al., 2000), and both suggest that there is a processing cost for recognizing that nonmatching cards need to be selected.

In sum, whichever approach is taken, there is either a processing benefit for matching cards or a processing cost for nonmatching cards. Of course, in terms of parameterization, these amount to the same thing. So we could parameterize the introduction of negations as either a small cost for nonmatching cards or a small benefit for matching cards. We opted for the former approach, simply because it is the one we favor, but it does not make any difference to the model-fitting exercise. Although this cost is presumed to arise as the result of cognitive processing, we can quantify its effects simply as a small reduction in the probability of selecting a nonmatching card, P_N. So, for example, the probability that each card should be chosen for the *if A then 2* rule can then be calculated as follows:

$$\begin{aligned} P(\text{TA-card}) &= 1 \\ P(\text{FA-card}) &= P_{\mathrm{F}}(1 - P_{\mathrm{C}}) - P_N \\ P(\text{TC-card}) &= 1 - P_{\mathrm{C}} \\ P(\text{FC-card}) &= P_{\mathrm{F}} - P_N. \end{aligned} \qquad (18)$$

And for the *if not-A then not-2* rule, these probabilities can be calculated as follows:

$$\begin{aligned} P(\text{TA-card}) &= 1 - P_N \\ P(\text{FA-card}) &= P_{\mathrm{F}}(1 - P_{\mathrm{C}}) \\ P(\text{TC-card}) &= 1 - P_{\mathrm{C}} - P_N \\ P(\text{FC-card}) &= P_{\mathrm{F}}, \end{aligned} \qquad (19)$$

and similarly for the other two rules. As for the information gain model, this means that parameterizing the mental models approach for the negations paradigm involves adding only one further parameter. To solve the problem created by the nonuniversal acceptance of the TA card (even when it matches), rather than 1, this value was set globally to .858, since this was the value that provided the best overall fit across all six experiments. Consequently, there were three free parameters in this model.

The results of the model fits are shown in Table A3 in the Appendix. The model could be rejected for only one of the six studies: Evans and Lynch (1973). The average $G^2(13) = 22.99$, $p > .02$. However, as for the information gain model, overall the model could be rejected [$G^2(78) = 137.94$, $p < .01$].

Discussion. Neither model avoided rejection overall. However, there are some points that suggest that the information gain model may be in better shape. First, to keep the playing field level, we have not attempted to compensate for the anomalous HL rule. If the exceptions parameter, ε, were high for this rule, this could allow for better fits. Indeed, an HL rule can make sense only if there are many exceptions (see Chater & Oaksford, 1999a). Of course, this would involve introducing more free parameters into the information gain model. However, at least there is an obvious and theoretically motivated way to proceed to improve the fit. Other than allowing the probability of selecting the TA card to vary, which has no theoretical motivation, there seems to be no obvious way in which to improve the fit of the mental models approach.

Second, even without adding further free parameters, there is a feature of these model fits that suggests that the information gain model might be doing a better job. It would appear that the parameters of the information gain model are much more stable across studies. That is, to capture the data as well as mental models, the parameters of the information gain model show much less between-study variation. To demonstrate this, we conducted F ratio tests of the homogeneity of variance (F ratio = 1) for all possible pairwise comparisons ($3 \times 3 = 9$) between the parameter values, with study as the unit of analysis. The ratio was computed with the variance for the mental models' parameters as the denominator. The mean F ratio across the nine comparisons was $F(5,5) = 0.27$ ($SD = 0.18$). That is, on average, the variance in the mental models' parameters was almost four times higher than the variance in the information gain model's parameters. Treating each comparison as an independent test of homogeneity of variance and combining the prob-

abilities, using Fisher's combined test (see above), this difference was highly significant [$\chi^2(18) = 39.67$, $p < .005$]. Thus, to explain the data in the mental models theory, it must be assumed that the probabilities that a conditional or a biconditional interpretation will be adopted or that a mental model will be fleshed out vary considerably between studies—certainly, much more so than the probabilities of the antecedent and the consequent in the information gain model. This large variation between studies seems hard to justify theoretically within the mental models or mental logic frameworks.

Summary. In this section, we first fitted the information gain model to each rule in the negations paradigm selection tasks analyzed by Oaksford and Chater (1994). The fits were very good. However, to compare the information gain model with the mental models theory required fitting these models to all 16 data points in a negations paradigm experiment. When this was done, these models appeared to provide comparable fits. However, although it was clear what direction to go in to improve the fit for the information gain model, it was not clear how to achieve the same goal for the mental models theory. Moreover, there was significantly more between-study variation in the mental models' parameters than in the information gain theory's parameters, variation that seems hard to justify theoretically.

TESTING THE NOVEL PREDICTIONS OF OPTIMAL DATA SELECTION

In the previous section, we concentrated on the data for which mental models/mental logic theories and the information gain theory all offer explanations. Our purpose was to show that the information gain model provides a better explanation of these data than do these other theories. Although explaining existing data better than other theories is desirable in a new theory, making novel confirmed predictions is also a desirable property. In this respect, the main novel prediction of the information gain model is that probability manipulations should affect selection task performance. Indeed, according to the contrast class account (Oaksford & Stenning, 1992), using high-probability categories in the antecedent and the consequent should produce effects similar to the varying of negations. Neither mental models nor mental logic theories make this prediction. Consequently, it makes no sense to provide detailed model comparisons. So, although we report the results of fitting the model to the data, we do not report the detailed fits in the Appendix. In this section, we will also briefly review work on the selection task that, although not directly testing the predictions of the information gain model, have produced results that are claimed not to be consistent with it.

Probabilities of Fictional Outcomes

Kirby (1994) developed a signal detection model of the selection task that predicted that as the probability of the antecedent, $P(p)$, increased, so the likelihood that participants would select the *not-q* card would increase. His model, however, failed to predict the systematic changes in selections of the other cards that occurred in his data. As $P(p)$ was increased, *not-p* card selections increased, and p card selections decreased (in Experiment 1, there was some evidence that q card selections also decreased, but this was not replicated in Experiments 2 and 3). The revised information gain model can account for these results. In these experiments, only $P(p)$ was manipulated. However, it is a constraint on the information gain model that the probability of the consequent must be greater than the probability of the antecedent weighted by 1 minus the probability of exceptions—that is, $P(q) > P(p)(1 - \varepsilon)$. When ε is low, as in these experiments, in which participants were told that ε is low (.01 or .1), this means that $P(q) \approx P(p)$. Figure 1 shows that respecting this constraint means that increasing $P(p)$ will increase the probability that *not-p* and *not-q* should be selected and will decrease the probability that p and q should be selected. That is, the information gain model is consistent with the observed pattern of effects. This is important because, according to Kirby's model, people are trying to detect falsifying p and *not-q* instances. However, these instances are not available for the *not-p* or the q card, and so Kirby's model cannot account for the pattern of results in his data.

As for the negations paradigm, we fitted the model to the data from each of the eight conditions in Kirby's (1994) Experiments 1–3. In Kirby's Experiment 1, there were two conditions, a small and a large $P(p)$ condition. In his Experiments 2 and 3, there were three conditions, a small, a medium, and a large $P(p)$ condition. For six out of the eight conditions, the model could not be rejected [mean $G^2(2) = 4.36$, $SD = 1.29$, $p > .10$]. Across these six conditions, the model could also not be rejected at the 1% level [$G^2(12) = 26.18$]. By experiment, the model could not be rejected for Experiment 1 [$G^2(4) = 7.36$, $p > .10$] or for Experiment 3 [$G^2(6) = 14.26$, $p > .02$]. Both conditions for which the model could be rejected were in Experiment 2—the medium and the large conditions. In both cases, the saturated model provided a significantly better fit to the data.

It would be premature to reject the information gain model on the basis of these two failures to fit the data, for several reasons. First, we will report other results using probabilistic manipulations that reveal good fits. Second, later on, we will discuss the kind of probability manipulation that would be expected to move people away from their default rarity values. Moreover, we will describe a recent experiment using an alternative and much more effective manipulation. Finally, only the information gain model or other probabilistic accounts (see below) predict the observed pattern of effects in Kirby's (1994) data.

It is also important that the best-fitting parameter values follow expectation—that is, they are low when they should be low and high when they should be high. Generally, this is the case when fitting the information gain model. However, when probabilities are manipulated,

the question arises as to the relationship between the manipulated probabilities (or *given* probabilities) and the best-fitting parameter values. What we have found in modeling Kirby's (1994) data and in the other results we report in this section is that this relationship follows Kahneman and Tversky's (1979) π function relating given probabilities to subjective probabilities. That is, people's subjective probabilities (best-fitting values) overestimate low given probabilities but underestimate high given probabilities. We will not discuss this issue further in reporting the results of experiments manipulating probabilities. This is because, later on, we will report the results of an experiment that seems to confirm that the subjective probabilities used to compute information gain do seem related to given probabilities as a π function.

The Reduced Array Selection Task

In the reduced array selection task (RAST), participants choose only between the *q* and the *not-q* options (hence, "reduced array"), and moreover, they are given the opportunity to see the data. For example, they might be told that they must test the rule that *all the circles are black* by picking out shapes from two boxes, one labeled "black shapes" and the other labeled "white shapes." Participants typically select shapes from both boxes, but on average, they select more shapes from the box containing white shapes. That is, far more falsificatory responding is observed (Johnson-Laird & Wason, 1970; Wason & Green, 1984). Oaksford and Chater (1994) argued that this is because the RAST makes explicit that the rule applies to a limited domain of cards and participants are told that there are equal numbers of *q* and *not-q* instances. It follows that the probability of the consequent, $P(q)$, is .5, violating the rarity assumption. When rarity is violated, the information gain of the *not-q* card is higher than that of the *q* card (Figure 1), and hence, the revised information gain model predicts more *not-q* card selections than *q* card selections.

Oaksford, Chater, Grainger, and Larkin (1997) tested this explanation of the RAST by systematically varying the probability of the consequent, $P(q)$. They used stacks of cards depicting colored shapes on one side, rather than boxes of colored shapes. The numbers of cards in each stack was varied to achieve the probability manipulation. By varying these probabilities, Oaksford et al. (1997) showed that the proportions of *q* and *not-q* cards selected varied in accordance with the information gain model—that is, as $P(q)$ fell, *q* card selections rose, and *not-q* card selections fell.

Figure 3 shows the results of Oaksford et al.'s (1997) Experiment 1. As can be seen, trends for the *q* and the *not-q* cards as the probability of the consequent, $P(q)$, rose were in line with the predictions of the information gain model. As $P(q)$ rose, there was a significant increase in the proportion of *not-q* cards selected, and there was a significant decrease in the proportion of *q* cards selected.

A possible alternative explanation for these effects is that the participants were selecting cards from the smallest stack or were selecting cards at random. In the low $P(q)$ condition, the smallest stack corresponded to the *q* card, and in the high $P(q)$ condition, the smallest stack corresponded to the *not-q* card. Consequently, a small stack bias could explain the pattern of selections in Oaks-

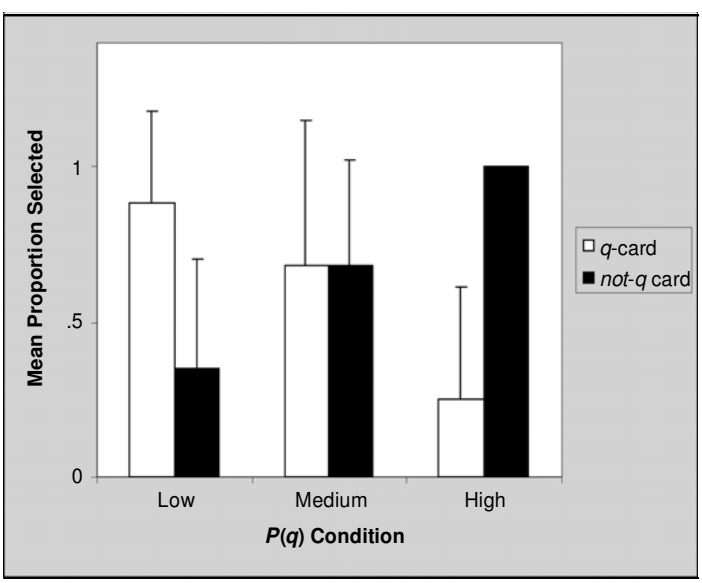

Figure 3. The results of Oaksford, Chater, Grainger, and Larkin's (1997) Experiment 1. The white bars indicate the mean proportion of *q* cards selected in each condition, and the black bars indicate the mean proportion of *not-q* cards selected in each condition. The error bars show a single standard deviation.

ford et al.'s (1997) Experiment 1. Oaksford et al. (1997, Experiment 3) therefore repeated that experiment, but now the participants selected cards from equal-sized stacks of cards. This was achieved by having the experimenter deal 10 cards from different sized packs, so although the probability information was available, the stack sizes were the same. The results of this experiment replicated their Experiment 1, confirming that the effects were indeed due to the probability manipulation.

Probabilities and the Standard Selection Task

Oaksford, Chater, and Grainger (1999) conducted a series of four experiments on the original four-card task, systematically varying the probabilities of the antecedent and the consequent of the conditional rule. According to the information gain model, if high- and low-probability categories are varied systematically between the antecedent and the consequent, the high-probability categories should produce results very similar to negations in the negations paradigm (see the Model Fitting II: The Negations Paradigm section). This is exactly what they observed.

We fitted the revised information gain model to each of Oaksford et al.'s (1999) experiments. In their Experiments 1 and 2, participants' belief in the rule [$P(M_D)$, high or low] was also manipulated. In Experiment 3 an *effects* manipulation (high or low) was included. Cognitive *effects* manipulations were proposed by Sperber, Cara, & Girotto (1995) to overcome habitual responses in the selection task by emphasizing the falsificatory *p*, *not-q* cases. Given these additional manipulations, there were 8 conditions in each of Oaksford et al.'s (1999) Experiments 1–3. There were 4 in Experiment 4. The model could not be rejected for any of the 28 conditions in these experiments [mean $G^2(2) = 3.15$, $SD = 2.31$, $p > .20$]. To assess the model across all the conditions, the 4 high-effects conditions in Experiment 3 were removed because Oaksford et al. (1999) argued that these were better explained by other probabilistic models that make explicit appeal to utilities (see below). Overall, the model could also not be rejected [$G^2(48) = 65.45$, $p > .02$]. By experiment, the model could not be rejected for Experiment 1 [$G^2(16) = 24.84$, $p > .05$], for Experiment 2 [$G^2(16) = 22.67$, $p > .10$], for Experiment 3 [$G^2(8) = 11.35$, $p > .10$], or for Experiment 4 [$G^2(8) = 6.69$, $p > .20$]. In sum, the model provided good fits to these data.

However, as Oaksford et al. (1999) conceded, their attempts to manipulate probabilities were not entirely successful. Experiments 1 and 2 used real-world contents—for example, *if a person is a politician then they are privately educated*—that were pretested for probability of occurrence. This raised the possibility that these materials may cue other relevant prior knowledge. In Experiment 1, although all the trends were in the right direction, the frequency of *not-q* card selections never exceeded that of *q* card selections. Oaksford et al. (1999) argued that this was because these materials did not provide a sufficiently powerful manipulation to overcome the default rarity assumption in data selection.

In Experiment 2 the high-probability antecedent rules reversed roles. That is, the HL rule produced results like the HH rule, and vice versa. Following other researchers in the area (Green & Over, 1998; Green, Over, & Pyne, 1997; Over & Jessop, 1998), Oaksford et al. (1999) suggested that this might be because participants were comparing the dependence model against different foils—that is, other than an independence model. To achieve the probability manipulation, they used rules such as *if an MP is a Conservative then he or she votes Labour in the general election*, which was an unbelievable, high-probability antecedent [high $P(p)$] and high-probability consequent [high $P(q)$] rule. The rule is unbelievable because it violates the strong belief that Conservative MPs vote anything but Labour in the general election. When Oaksford et al. (1999) used the opposite to the dependence model as a foil hypothesis instead of the independence model (i.e., the dependency is between being a Conservative MP and *not* voting Labour), they found good fits to the data. It is the fit using this foil model in the revised information gain account that we reported above. To avoid the effects of prior beliefs like this, Oaksford et al. (1999) used abstract material in their Experiments 3 and 4 and found results much more in line with predictions.

Causal Selection Tasks

Further evidence consistent with the revised information gain model has recently been presented by Green and Over (1997, 2000; see also Over & Jessop, 1998). Green and Over (1997) tested the Bayesian account of data selection by having participants test the causal relation, *if a person has Zav's disease, then they have a raised temperature.* They would be asked, for example, "how many out of 100 patients already diagnosed with Zav's disease do you want to take the temperature of?" (*p* card selections). This manipulation provided data for the first time on within-subjects selection tendencies. According to mental models (Johnson-Laird & Byrne, 1991), only four selection patterns are possible (see the Mental Logic and Mental Models section above). Recall that if models for the conditional or the biconditional interpretation are not fleshed out, participants must select either only the *p* card or the *p* and the *q* cards, respectively. If these models are fleshed out, then on the conditional interpretation, participants should select the *p* and the *not-q* cards, and on the biconditional interpretation, they should select all four cards. None of these interpretations is consistent with Green and Over's (1997) results. Around 70% of the participants wanted to examine some of all four classes of patients but wanted to see more patients that corresponded to the *p* and *q* cards than to the *not-p* and *not-q* cards. As Green and Over (1997) observed, this finding is consistent only with Bayesian probabilistic accounts, such as information gain.

Green and Over's (1997) response procedure allows participants to reveal the underlying probabilistic basis of their selection decisions. These are continuous data, and so we cannot really model them in the same way as we have until now. However, we can think of each partici-

pant's response to each card—that is, "*x* out of 100"—as a response frequency, as if they had experienced 100 trials. This, of course, vastly inflates the value of *N*, the number of responses, which increases the probability of a poor fit (Read & Cressie, 1988). Nonetheless, when fitted to Green and Over's (1997) Experiment 1, mild condition data (see below), the revised information gain model could not be rejected at the 1% level [$G^2(2) = 7.61$, $p > .02$]. If people are generating an analogue of the probabilities calculated in the revised information gain model, it is straightforward to convert these to frequencies of patients they wish to look at. It is much harder to envisage how a consistent logical interpretation of the conditional, as embodied in mental logic and mental models theories, could account for these within-subjects preferences.

Green and Over (1997) also manipulated the utility of finding out whether a raised temperature is diagnostic of Zav's disease by providing information about whether the disease is life threatening (serious condition) or not (mild condition). The information gain model does not incorporate utilities, so we did not attempt to model this manipulation (see the Other Probabilistic Accounts of Data Selection section below).

Green and Over (2000) used a scenario very similar to that in Green and Over (1997), but with cholera as the disease. However, the participants were asked only whether they wanted to see, for example, "villagers already diagnosed as having cholera" (*p* card), rather than how many they wanted to see. Consequently, the task yielded binary response data, as in the standard selection task. The rule used was, *if you drink from the well then you will get cholera*. The main experimental manipulation involved telling participants either that *most* villagers have cholera and *most* drink from the well, which corresponds to a high-probability antecedent [high $P(p)$] and high-probability consequent [high $P(q)$] condition, or that *few* villagers have cholera and *few* drink from the well, which corresponds to a low $P(p)$ and a low $P(q)$ condition. The fit of the revised information gain model to Green and Over's (2000) data (category condition) was very good: In the *few* condition, $G^2(2) = .88$, $p > .20$, and in the *most* condition, $G^2(2) = .14$, $p > .20$. In the *few* condition, $P(p) = .42$ and $P(q) = .40$, and in the *most* condition, $P(p) = .61$ and $P(q) = .59$. In both cases, $P(p) \approx P(q)$. According to our model, this entails that the participants were treating the dependence model as a biconditional (i.e., most of the probability is located in the *p*,*q* and the *not-p*, *not-q* cells; indeed, if $P(p) = P(q)$ and $\varepsilon = 0$, all the probability is located in these two cells). That is, they were testing a model in which drinking well water was both necessary and sufficient for catching cholera. This may be a feature of causal selection tasks, in which the rule describes a putative causal regularity.

The results reported in this section seem to support the view that "no account of the selection task is sufficiently general if it cannot take account of the set size of *p* and the set size of *q* or the probability judgments which reflect these" (Green & Over, 2000, p. 66). That is, any explanation of the selection task must take a probabilistic approach as embodied in the information gain model. However, there have been some apparent failures to replicate these probabilistic effects.

Probabilities or Coherence Bias?

Oberauer et al. (1999) carried out three experiments that, they argued, all failed to replicate the effects we reviewed above. These findings led them to the conclusion that "optimal data selection does not explain the selection task" (Oberauer et al., 1999, p. 141). It is difficult to interpret such failures to replicate. However, in this case, the failure was only partial. In their Experiment 1, they found trends that were consistent with optimal data selection. Oberauer et al.'s main argument hinges on the poor fits they obtained using the values of $P(p)$ and $P(q)$ that participants were given experimentally. However, as we mentioned in the Probabilities of Fictional Outcomes section, it is unlikely that experimental manipulations will affect people's subjective probabilities so directly (see also Evans & Over, 1996a; Hattori, 2002; McKenzie, 2000; McKenzie, Ferreira, Mikkelsen, McDermott, & Skrable, 2001; McKenzie & Mikkelsen, 2000).

We therefore fitted the model to the data from Oberauer et al.'s (1999) Experiment 1 in the same way as in the rest of this paper: by seeking the parameter values that provided the best fit. The fit of the model to the data was comparable, if not better, than the fits we have already reported [mean $G^2(2) = 0.84$, $p > .20$], across all four conditions [$G^2(8) = 3.38$, $p > .20$]. Moreover, when the parameters were supposed to be high ($M = .44$, $SD = .02$), they were higher than when they were supposed to be low [$M = .28$, $SD = .12$; $t(5) = 2.88$, $p < .025$]. Consequently, it would seem that, contrary to Oberauer et al., the revised information gain model can provide very good fits to their data.

How did Oberauer et al. (1999) explain the results of their Experiment 1, which as we have shown, would appear to provide good evidence for the information gain model? They suggested that the categories they used—for example, numbers between 1 and 10 or between 10 and 1,000—lack coherence in that they do not all share some common property. Oberauer et al. therefore suggested that lacking such a coherent basis, a more coherent category will be one with fewer members. They then argued that people were demonstrating a *coherence bias*—that is, they were selecting the cards that corresponded to a lower prior probability. They argued that if coherence were to be restored, the putative probability effects observed in their Experiment 1 would disappear. They achieved this manipulation by using categories such as vowel and consonant in their Experiment 2 and A (or B) and 1 (or 2) in their Experiment 3. Probabilities were manipulated by indicating that a certain number of cards had vowels, or As, and so forth, on them. Neither experiment revealed any effects of the probability manipulation.

However, Oaksford et al. (1999) used coherent real-world categories in their Experiments 1 and 2 and co-

herent abstract materials in their Experiment 4 and observed many of the probabilistic effects predicted by optimal data selection accounts. Consequently, coherence bias is unlikely to be the explanation of the results of Oberauer et al.'s (1999) Experiment 1. Moreover, there is a factor that was present in Oberauer et al.'s Experiments 2 and 3 that was not present in their Experiment 1 or in Oaksford et al.'s experiments. Oberauer et al.'s Experiments 2 and 3 involved two sample selection phases. First, the participants were given probability information about a large pack of cards (1,000) from which, they were told, a smaller sample had been selected at random (50 or 52). Second, the four selection task cards were then drawn from this smaller sample. Consequently, for this manipulation to work, it must be assumed that the participants treated the smaller random sample as *representative* of the larger pack of cards from which it had been drawn. However, the participants' understanding of the probability manipulation was assessed only with respect to the larger pack, and not the smaller sample. Therefore, there is no evidence that Oberauer et al.'s participants treated the smaller sample as representative of the larger pack. Indeed, to preserve the known probability distribution of the larger pack in the sample, representative, as opposed to random, sampling is required. Because of the distribution in the larger packs, in Oberauer et al.'s Experiments 2 and 3, it was possible for $P(p)$ and $P(q)$ (in the sample) to take on any value between 0 and 1. Given this uncertainty, we would argue that people have simply made the default rarity assumption for all four rules. This hypothesis has been tested recently by Oaksford and Wakefield (2003), who used a natural sampling paradigm to manipulate probabilities.

Probabilities and Natural Sampling

Recently, Oaksford and Wakefield (2003) conducted an experiment using the same materials as those in Oberauer et al.'s (1999) Experiment 3, but without the second sample selection phase. The main purpose of this experiment was to test two hypotheses. First, Oaksford and Wakefield argued that providing probability information via *natural sampling* (Gigerenzer & Hoffrage, 1995) should lead to a stronger manipulation. According to Gigerenzer and Hoffrage, manipulating probability information experimentally is best achieved by manipulations that simulate the way people normally acquire this information in their natural environment—that is, by experiencing the instances used to compute a relative frequency one at a time. For example, people compute the probability of a bird's being black by observing individual birds and storing information about sample size and the number of black birds. Such *natural sampling* was implemented in Oaksford and Wakefield's experiment by showing participants 40 cards, one at a time. The proportion of *p*, *not-p*, *q*, and *not-q* cards reflected the probability information the participants were given before performing the selection task. Second, Oaksford and Wakefield argued that obtaining probability estimates from participants indirectly should more accurately reflect the values they use in data selection. Therefore, after the selection task, the participants were asked to classify 50 cards "drawn" from the pack into the four possible card types: *p* and *q*, *p* and *not-q*, *not-p* and *q*, and *not-p* and *not-q*. The proportions of each card type provided estimates of the four cells in the joint probability distribution for *p* and *q* (i.e., the upper half of Table 1), from which all the parameters of the model could be calculated.

The results (see Figure 4) showed all the probability effects predicted by the information gain model. This was impressive, given that the materials were identical to those used by Oberauer et al. (1999), who failed to find any effects from manipulating probabilities. There were also some other interesting effects observed. First, the indirect estimates of the parameters of the model overestimated the low probabilities and underestimated the high probabilities given in the experimental set-up. This directly mirrors a π function relating the experimentally given probabilities and the best-fitting estimates (see the Probabilities of Fictional Outcomes section). Importantly, the best-fitting parameter values were higher when they were predicted to be high than when they were predicted to be low [$t(6) = 5.10$, $p < .0025$], and they were highly significantly correlated with the given probabilities [$r(6) = .89$, $p < .005$].

Second, when the indirect estimates of the parameter values were weighted for prior knowledge of rarity (by averaging with the indirect estimates for the LL rule), they provided good fits to the data. When these values were used, the model could not be rejected for any condition [mean $G^2(4) = 7.18$, $p > .10$], and it could not be rejected across conditions [$G^2(16) = 28.70$, $p > .02$]. Oaksford and Wakefield (2003) also showed that the information gain model provided better fits than a post hoc model of probability effects proposed by Oberauer et al. (1999) to explain the results of their Experiment 1. That is, contrary to Oberauer et al., experimentally derived values of the parameters of the information gain model can show good fits to the data. However, indirect estimates (reflecting a π function) corrected for the effects of prior knowledge must be used. Consequently, contrary to Oberauer et al.'s conclusion, optimal data selection does explain the selection task.

Results Not Consistent With the Information Gain Model

We now will review some results on data selection that, it is claimed, are not consistent with the information gain model. Most of these results involve *facilitating* the logical response without manipulating probabilities.

Group reasoning. In a single experiment, Moshman and Geil (1998) showed that solving the task in groups leads to higher solution rates. This result is very interesting but stands in need of replication and further study. An aspect of the results that is interesting is that, in nearly all the groups, at least one person initially chose the *not-q* card. Consequently, there was always someone who needed to account for why they had done this. The

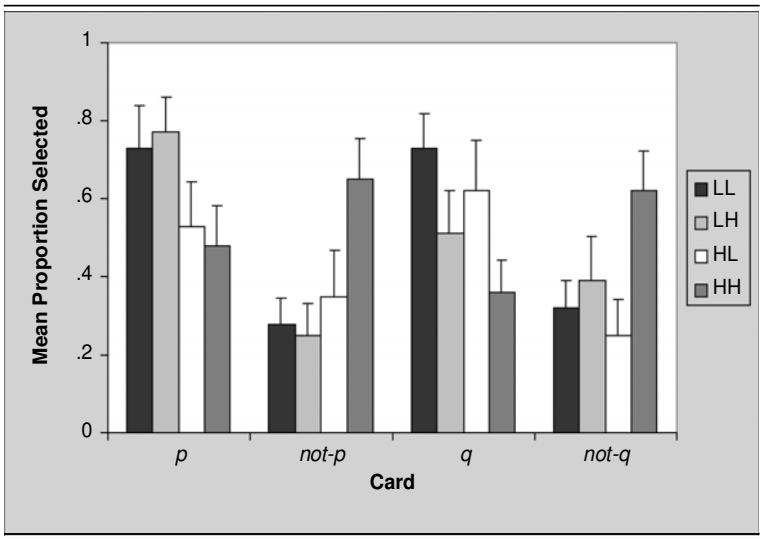

Figure 4. The mean proportion of cards selected in each condition in Oaksford and Wakefield (2003), with error bars showing standard error. LL, low-probability antecedent, low-probability consequent; LH, low-probability antecedent, high-probability consequent; HL, high-probability antecedent, low-probability consequent; HH, high-probability antecedent, high-probability consequent.

question is what changes people's minds. One answer is that they develop true logical insight into the falsificationist nature of hypothesis testing (Moshman & Geil, 1998). As we have argued elsewhere (Oaksford & Chater, 1994), since it is philosophically debatable whether hypothesis testing is best regarded as logical (e.g., Howson & Urbach, 1989), this is a dubious argument. It seems more likely that people come to adopt an interpretation of the rule in which *p*, *not-q* instances definitely make the rule false and prior knowledge is irrelevant. The former happens even in our probabilistic model when $\varepsilon = 0$. The latter occurs when people realize that the rule applies only to the four cards. Under this interpretation, prior knowledge, which affects the probabilities of the antecedent and the consequent, is irrelevant. What would be interesting is to probe the participants' understanding of the rule before and after the group discussion, rather than indexing such changes simply on the basis of changes in response patterns on the selection task.

Getting the right interpretation. Gebauer and Laming (1997; see also Osman & Laming, 2001) argued that performance in their experiments was logical when participants' interpretations of the task rule were taken into account. In their Experiment 1, although most participants failed to give the logical response, it seemed that this was because they misinterpreted the rule. For example, *one side . . . other side* may be conflated with *front . . . back*. When these interpretations were taken into account, performance could be interpreted as logical. This interesting result it is not consistent with the attempts to encourage a logical interpretation in the early literature on the selection task, which focused on exactly the same possible misinterpretations (Wason & Johnson-Laird, 1970, 1972). These earlier attempts to remove these possible misinterpretations completely failed to improve performance. Gebauer and Laming (1997; like Osman & Laming, 2001) did not attempt to explain this apparent inconsistency.

Moreover, if they are right, their findings do not just threaten probabilistic accounts. Gebauer and Laming (1997) argued that their misinterpretation account is consistent with formal rule theories and mental models. However, according to these accounts, the rules are not misinterpreted in the way that *one side . . . other side* may be conflated with *front . . . back*. Rather, the rules are interpreted as either conditionals or biconditionals *assuming the "one side . . . other side" interpretation*. However, in mental models theory, people normally represent only part of the meaning of the original conditional—that is, one line of the corresponding truth table. It is this *partial* interpretation that explains performance, not any *mis*interpretation of the rules. This is absolutely clear from the fact that this partial interpretation is taken to apply to all conditionals, whether they appear in selection tasks that introduce *one side . . . other side* ambiguities or not. Consequently, psycho-logic accounts, such as mental models and mental logic, are as much at risk if Gebauer and Laming are correct as probabilistic accounts. Thus, although at present we offer no explanation of these findings, it should be borne in mind that they potentially invalidate everyone's explanation of these data. Much more research is therefore clearly going to be required before anyone is going to be convinced by these results.

Facilitation without probabilities. Almor and Sloman (1996) presented evidence that certain conditional

rules reliably produced the logical response although probabilities were not manipulated. However, Oaksford and Chater (1996) argued that these rules fell into two categories that meant that the information gain model did not apply. First, some of the rules were analytic truths—for example, *if a large object is stored then a large container must be used*—for which evidence is irrelevant. It is part of the meaning of *large object* that if it is stored, a large container must be used. Since this has to be true, there is no uncertainty to be reduced. Second, the remaining rules were deontic regulations: *If a product gets a prestigious prize then it must have a distinctive quality.* Oaksford and Chater (1994) dealt with deontic regulations in a separate theory in which people were argued to maximize expected utility rather than information gain. Consequently, these apparent displays of logicality do not impugn the claim that when participants construe their task as selecting data to test a hypothesis, they seek to maximize information gain.

Varying the card array. Hardman (1998) introduced an interesting manipulation in which certain cards were removed from the four-card array and substituted by an additional copy of one of the other cards. He argued that this manipulation should affect the predictions of the information gain model because card choice is competitive—that is, card informativeness is scaled by the total amount of information available (see Equation 7). Thus, if a highly informative card (e.g., *p*) is replaced by an uninformative card (e.g., *not-p*), the informativeness of the two *not-p* cards should rise, since they now represent a greater proportion of the information available. Hardman presented three experiments in which different replacement strategies and different rules were used, either a standard affirmative rule (*if p then q*) or a negated consequent rule (*if p then not-q*). In each experiment, he argued that his results were not consistent with the information gain model. For example, suppose that the *p* card has an information gain of .45 bits, the *q* card .35 bits, the *not-p* card .05 bits, and the *not-q* card .15 bits. Then, according to the scaling procedure with the standard <*p*, *not-p*, *q*, *not-q*> card array, the scaled informativeness values would be unchanged at <.45, .05, .35, .15>. But given the array <*not-p*, *not-p*, *not-p*, *not-q*>, the scaled informativeness values would be <.17, .17, .17, .50>, which predicts that there should be considerable increases in *not-p* and *not-q* card selections, as compared with the standard array. Hardman did not observe these predicted changes in card selections. For example, in his Experiment 3, although he observed the change for the *not-q* card in our example, he observed no similar changes for the *not-p* cards.

As Hardman (1998) observed, this problem does not arise if the information gains are not rescaled according to Equation 7. We originally scaled the information gain in this way by analogy to foraging models of food patch selection (Myerson & Miezen, 1980; Oaksford & Chater, 1998b; Pirolli & Card, 1999), where animals disperse their foraging activities between food patches, depending on the total food available. Of course, before deciding how to disperse its time between patches, an animal must decide which patches it is worth dispersing its time between. Suppose this decision is made pairwise by first comparing the amount of food in the largest patch with the next largest: If the ratio of the smaller to the larger patch is greater than .5, then forage at this smaller site as well, and so on. If we do this calculation for the *informavores* (Dennett, 1991) in the last paragraph, then for the standard card array, the process will stop at the *q* card, which is more than half as informative as the *p* card but is more than twice as informative as the *not-q* card. However, in the <*not-p*, *not-p*, *not-p*, *not-q*> array, the *not-q* card is more than twice as informative as any individual *not-p* card, so the latter are not chosen, which may explain Hardman's results. This process determines the number of cards selected before rescaling to determine the strength of conviction that the card should be turned, as was suggested by Chater and Oaksford (1999a).

Summary

The revised information gain model provided detailed fits to the data explained by the original model (Oaksford & Chater, 1994). Moreover, six studies (Green & Over, 1997, 2000; Oaksford et al., 1999; Oaksford et al., 1997; Oaksford & Wakefield, 2003; Oberauer et al., 1999, Experiment 1) and a total of 14 experiments have produced results consistent with the revised information gain model. We also showed how the revised model provided good fits to 10 of these experiments. Showing that the model can provide good fits to these new data sets is important because some of the parameters of the revised model were estimated against some of the data originally modeled by Oaksford and Chater (1994). Specifically, the parameters of the selection tendency function (Hattori, 1999) were estimated in this way. However, to model these new data sets, we retained the values of these parameters [and of $P(\mathrm{M_I})$ and ε] that we used to model the original data sets, and we still found good fits by simply allowing $P(p)$ and $P(q)$ to vary. The best-fitting parameters were invariably interpretable in the way we argued they should be—that is, they were high when they should be high and low when they should be low. Moreover, when tested indirectly, people's subjective probability values mirrored the best-fitting parameter values and were related to given probabilities as a π function. We have also presented arguments showing that many results argued to be not consistent with information gain do not discriminate against the model to anything like the degree claimed.

THEORETICAL DISCUSSION

In this closing section of the paper, we will discuss alternative theoretical proposals within a probabilistic framework and will reply to a variety of theoretical objections to the information gain model that have appeared in the literature. In addressing these theoretical objections, we will concentrate solely on those that have arisen in the literature but that we have not addressed

elsewhere before (i.e., in Oaksford & Chater, 1996, 1998a, 1998b, Oaksford et al., 1999, or Oaksford et al., 1997).

Other Probabilistic Approaches to Data Selection

There are now a variety of other probabilistic accounts of the selection task (Evans & Over, 1996a, 1996b; Klauer, 1999; Nickerson, 1996; Over & Evans, 1994; Over & Jessop, 1998). Oaksford et al. (1999) reviewed these accounts, which can all be encompassed within the optimal experimental design approach (Berger, 1985; Fedorov, 1972) and, consequently, all share the basic underlying structure of the information gain model. The main theoretical difference is how each model formalizes the notion of informativeness. The critical difference concerns whether a *disinterested observer* or a *decision-theoretic* approach is taken to inquiry (Chater, Crocker, & Pickering, 1998; Chater & Oaksford, 1999a). On the disinterested observer approach (Nickerson, 1996; Oaksford & Chater, 1994), it is assumed that participants are not biased toward any particular type of evidence by their current goals. Consequently, their data selection behavior is influenced only by the relevant probabilities. On the other hand, someone might be seeking evidence that, for example, drinking water from the well makes you ill. With this goal in mind, they will place greater value on finding evidence of someone's being ill after drinking the well water. This is because the costs of erroneously rejecting this hypothesis are very great: Many people will continue to get ill. This is an example of a decision-theoretic approach (Evans & Over, 1996a, 1996b; Klauer, 1999; Over & Evans, 1994; Over & Jessop, 1998). People are inquiring into their world with a particular decision problem in mind: Should you drink the well water or not?[2] The important costs and benefits relate to accepting or rejecting a hypothesis—that is, the Type I and Type II errors in standard hypothesis testing.

These other models point to a convergence of opinion that a probabilistic approach is the right way to explain the indicative and the causal selection task. However, there is a clear disagreement about whether a disinterested or a decision-theoretic framework should be adopted. We think that each is equally valid but that care must be taken about when each should be applied. It seems to us that unless the experimental set-up can provide clear-cut utilities for making Type I or Type II errors, a disinterested approach is clearly more appropriate. For example, Oaksford et al. (1999) argued that Sperber et al. (1995) *effects* manipulation can be explained better by decision-theoretic approaches, such as those in Evans and Over (1996b) and Klauer (1999). The effects manipulation involved making the *p*, *not-q* instance salient by creating a context in which it is diagnostic of a fault—for example, a machine that is supposed to be printing cards according to the rule that if there is a circle then the card is blue starts printing red circles. This manipulation can be regarded as raising the costs of failing to reject a hypothesis when it is false (i.e., failing to detect a fault). A decision-theoretic perspective may also be more appropriate for explaining some of the results of Green and Over (1997), who also manipulated the seriousness of an illness and, thereby, the costs associated with failing to detect the illness. Nonetheless, in most selection task experiments that use abstract material or use contentful indicative rules without a context introducing explicit utilities, a disinterested approach seems more appropriate.

However, these approaches are often represented as being in competition rather than as complementary (Evans, 1999; Evans & Over, 1996b; Green, 2000; Klauer, 1999). That is, it is argued that the decision-theoretic approach should be seen as an alternative and descriptively more adequate way of explaining the data. The key empirical issue concerns the effect of believability, where the disinterested and the decision-theoretic approaches diverge. Whereas the decision-theoretic approach predicts that if people disbelieve the rule they should select more *not-q* cards than *q* cards, disinterested approaches predict no effect of believability. Klauer cites several studies that seem to show results consistent with the decision-theoretic approach (Fiedler & Hertel, 1994; Love & Kessler, 1995; Pollard & Evans, 1981, 1983). However, Chater and Oaksford (1999a) argued that in all these studies, only exceptions were manipulated—that is, the incidence of *p*, *not-q* instances—and not believability per se. As they pointed out, it is possible to believe strongly a rule that has many exceptions. For example, many people believe quite strongly that allowing children to walk home from school increases their chance of being abducted, although the probability of being abducted while walking home from school is tiny. That is, the probability of exceptions and the degree of belief in a rule can be independent.

Only two studies have explicitly manipulated believability in the selection task: Green and Over (1997) and Oaksford et al. (1999). In their Experiment 2, Green and Over (1997) found that an almost identical proportion of participants turned the *not-q* card in the believed true (55.3%) and in the believed false (54.9%) conditions. Similar results were obtained in Oaksford et al.'s (1999) Experiment 1 (high-belief condition, 29.7% *not-q* card selections; low-belief condition, 25%) and in their Experiment 2 (high-belief condition, 33.1%; low-belief condition, 33.4%). Consequently, it would seem that the best interpretation of these models is that they apply in different situations. The challenge for proponents of the decision-theoretic approach is to demonstrate believability effects in contexts in which the utilities are well defined.

We have now seen that the information gain model seems to account for the existing data better than do other theoretical proposals. It also appears to be well supported by data confirming its novel predictions. It is also defendable in the light of data with which it is apparently inconsistent. Finally, as we have just seen, it provides better explanations than do closely related probabilistic theories. However, despite these apparent successes, there may be some underlying theoretical prob-

lems with the theory that still rule it out as somehow incoherent or unlikely to be psychologically real. The rest of this theoretical discussion addresses a variety of theoretical objections that have been proposed in the literature since the model first appeared.

Model Fits

Laming (1996) argued that the original optimal data selection model was able to provide only ordinal fits to the data—that is, we modeled only the rank order in participants' card selections. However, in this paper, we have shown that the optimal data selection model can provide excellent fits not only to the ordinal trend in the data, but also to the exact location.

Information Gain Versus Other Information Measures

Some critics have been concerned that we narrowly focused our account on only one particular measure of information, Shannon–Wiener information (Shannon & Weaver, 1949; Wiener, 1948), although there are other possibilities (Evans & Over, 1996b; Klauer, 1999; Laming, 1996; Oberauer et al., 1999). The review of other measures (see the Other Probabilistic Approaches to Data Selection section) shows a general feature of the optimal data selection models. That is, although they all propose different measures of informativeness, given the rarity assumption, they all make similar predictions (with the exception of the believability predictions we discussed above). That is, given the rarity assumption, the optimal data selection approach can quite robustly predict the main findings even under changes of the particular information measure used.

Sequential Sampling

Laming (1996) criticized our original model because a Bayesian account should involve sequential sampling—that is, participants should turn the most informative card, revise their priors, reassess the informativeness of the cards, pick the next most informative card, and so on. Of course, this is not what happens in the selection task, because participants never actually turn the cards over. The main point to make here (but see also Oaksford & Chater, 1998b) is that Klauer (1999) has modeled the selection task using both sequential and *nonsequential* Bayesian models. The predictions agreed with the information gain model. Consequently, a nonsequential Bayesian account not only makes sense (see Chater & Oaksford, 1999a), but also makes predictions similar to those of a sequential account.

Alternative Models

We suggested that participants compared the rule—that is, the dependence model—with an independence model. Various authors have criticized us for this choice on varying grounds (Green & Over, 1997, 1998; Green et al., 1997; Laming, 1996; Oberauer et al., 1999; Over & Jessop, 1998). First, we have been criticized for proposing that participants are testing a particular dependence model against just the possibility that there is no dependency between p and q, rather than against every other possible dependency between p and q (Laming, 1996; Oberauer et al., 1999). However, without any prior knowledge about other possible relationships between p and q, it seems psychologically plausible that the only alternative considered is no relationship. Moreover, the dependency between p and q that we always had in mind was *causal* sufficiency. In most recent models of causal judgment, the independence model is always the foil against which the presence of a causal dependency is assessed (e.g., Cheng & Novick, 1990). In these models, a causal dependency (however weak) is taken to exist between p and q if $P(q \mid p) > P(q)$ (positive causal relationship) or if $P(q \mid p) < P(q)$ (negative causal relationship). Thus, in attempting to construct a theory of data selection, the independence model seemed to be the most justified. The real question is the nature of the infinite number of possible causal dependencies that might exist between p and q. However, it turns out that the nature of the dependency between p and q and the data one should select are relatively independent. That is, variation in $P(q \mid p)$ $(1 - \varepsilon)$ has little effect on the data people should select. For example, it never affects the rank order of informativeness over the four cards when $P(p)$ and $P(q)$ are kept constant. Moreover, other prior knowledge constrains the relevant probabilities. First, the conditional statement *if p then q* clearly suggests a positive causal dependency, which indicates that ε is low, which is why we set it to .1 in all the model fits we report. Second, causal sufficiency suggests that $P(q) > P(p)$. Finally, the rarity assumption indicates that $P(p)$ and $P(q)$ are low.

Second, other authors have pointed out that prior knowledge may suggest better foil models or, indeed, more than one such model (Green & Over, 1997, 1998; Green et al., 1997; Over & Jessop, 1998). The information gain model can incorporate these possibilities. Indeed, Oaksford et al. (1999) invoked just such an alternative foil in modeling the results of their Experiment 2 (see above). Moreover, in contrast to the log-likelihood ratio (Evans & Over, 1996b), the information gain measure can incorporate many different hypotheses.

Exceptions

Evans and Over (1996b) argued that the original dependence model was also incapable of explaining data from Kirby (1994) or from Pollard and Evans (1983). Oaksford and Chater (1998b) argued that this was because the original dependence model did not allow exceptions. Kirby told his participants that a machine printing cards had made an error—that is, it produced a *p*, *not-q* instance. Consequently, an exceptionless generalization was already known to be impossible, and therefore, the independence model had to be true. Oaksford and Chater (1998b) conceded this problem with the original model and so modified it, as in the upper half of Table 1, to include an exceptions parameter. This new

model showed good fits to most of Kirby's data, as we have shown.[3]

Parameter Alignment

The original model has been criticized (Green & Over, 1997, 1998; Green, Over, & Pyne, 1997; Laming, 1996; McKenzie & Mikkelsen, 2000) on the grounds that the particular form of the dependence model was used only to explain the data. In particular, it has been argued that this model was used simply to guarantee that the *not-p* card, which is rarely selected, would be completely uninformative (Laming, 1996). This selection of models also had the odd consequence, noted by Green and Over (1998), that the probability of the consequent varied between models (b in M_I and $a + b(1 - a)$ in M_D). Green and Over (1997) also pointed out that, contrary to the predictions of the original model, the *not-p* card can provide useful information (see also McKenzie & Mikkelsen, 2000). All of these problems have been resolved in the dependence hypothesis used in the revised information gain model. In that model, the *not-p* card can be informative, the probability of the consequent is b in both models and yet, contrary to what one might expect from Laming's (1996) argument, the revised model still provided good fits to the data. Moreover, the very same dependence model has been shown to provide good fits to data sets in other areas of conditional reasoning (Oaksford et al., 2000).

Biconditional Interpretation

The revised model also resolves a problem raised by Oberauer et al. (1999) for Oaksford and Chater's (1994) account of Kirby's (1994) data, where probabilities were manipulated in an abstract selection task for the first time. To model these data, we assumed that $P(p) = P(q)$ (although Oaksford & Chater, 1998b, relaxed this assumption and still showed good fits to the data). Oberauer et al. objected to this assumption on the grounds that for the q and *not-q* cards, the participants were always told that either a "+" or a "−" was printed on a card. However, as Over and Evans (1994) pointed out, Kirby did not tell the participants that these symbols were printed at random, so this fact does not license any particular value for the probability of finding one of these symbols on a card. So, contrary to Oberauer et al.'s suggestion, it is certainly not incoherent to propose that these probabilities vary with $P(p)$.

Oberauer et al. (1999) also argued that we should have used a biconditional dependence model, because simply equating the probability of the antecedent and the consequent—that is, $P(p) = P(q)$—does not achieve this interpretation in the original dependence model. They introduced a biconditional model, where $P(p, \textit{not-q}) = P(\textit{not-p}, q) = 0$, and showed poor fits to Kirby's (1994) results. However, in the revised dependence model when $P(p) = P(q)$ and $\varepsilon = 0$, then $P(p, \textit{not-q}) = P(\textit{not-p}, q) = 0$—that is, in the revised model, the biconditional interpretation and the assumption that $P(p) = P(q)$ go hand in glove (see also Hattori, 2002). Of course, it was the revised dependence model that we used to model Kirby's data above, and we showed good fits to most of the data without making any of the assumptions that Oberauer et al. (1999) criticized.

Rarity Assumption

Some authors have objected to the rarity assumption (Laming, 1996; Oberauer et al., 1999). That is, to explain the data, it must be assumed that $P(p)$ and $P(q)$ are small. Recently, however, it has been shown that rarity is the default when people are testing (Anderson & Sheu, 1995; McKenzie & Mikkelsen, 2000) or framing (McKenzie et al., 2001) hypotheses. McKenzie and Mikkelsen showed that people regard rare evidence as more relevant to supporting a hypothesis than common evidence. So for example, logically, both black ravens and nonblack nonravens (e.g., pink flamingos) confirm the hypothesis that *if it's a raven then it's black*. However, people regard black ravens as more supportive of this hypothesis (see also Oaksford & Chater, 1996). Moreover, rare observations were often selected even when they were not mentioned in the hypothesis.

Perhaps people are simply matching the salient named items (see Evans, 1989). However, McKenzie et al. (2001) showed that hypotheses are normally phrased in terms of rare events, so that such a matching strategy is invariably the rational thing to do. They showed participants' data about a group of students' Scholastic Aptitude Test (SAT) scores and whether these students were admitted to a select university. Only one student was admitted, and this was the only student with a high SAT score. When asked to fill in a sentence frame describing this situation, "If ____, then ____," the participants strongly preferred the phrasing "if applicants have high SAT scores, they will be accepted" over "if applicants have low SAT scores, they will be rejected," even though both are equally legitimate ways to complete the statement. Crucially, when given the information that most students were accepted and that few applicants had low SAT scores, this finding was reversed—that is, they now preferred the second phrasing. In both cases, the participants preferred to frame a hypothesis in terms of rare, rather than common, events. In sum, not only does the rarity assumption make conceptual sense of the literature in the philosophy of science (see Mackie, 1963; Oaksford & Chater, 1996), it also is a part of people's normal expectations about the hypotheses they formulate and test about their everyday world.

An important observation made by McKenzie (2000) is that the rarity assumption "is presumed to exist because of lifelong learning that presence is rarer than absence" (p. C7). Consequently, experimental manipulations that violate rarity are unlikely to totally eliminate the tendency to select the p and q cards in the selection task (see Oaksford et al., 1999, and Oaksford et al., 1997, for a similar argument). This is because "violations of rarity move participants' behavior in the appropriate di-

rection but not by as much as the normative theory (unencumbered by strong priors about how the world usually works) would predict" (McKenzie, 2000, p. C7). This argument is consistent with our findings that although our probability manipulations had a large effect on the best-fitting parameter values, calibration was not perfect—that is, it followed a π function. That is, the observed behavior was consistent with the probability manipulation, but the changes were not as extreme as the model would predict if the experimentally given values were simply plugged into the model.

Revising $P(p)$

Several authors (Green & Over, 1998; Oberauer et al., 1999) have questioned the revision strategy for the *if not-p then q* rule [or assuming the contrast set account of negations, the high $P(p)$ and low $P(q)$ rule (HL)]. Indeed, Green and Over (1998) provide an example for which it would be incoherent to adopt this strategy. However, the example relies on allowing $P(q)$ to vary between models, which is no longer possible in the revised model (see the Parameter Alignment section). However, when there are exceptions—that is, $\varepsilon > 0$—the only constraint is that $P(p)(1 - \varepsilon) \leq P(q)$. Consequently, $P(p)$ can be greater than $P(q)$ as long as participants are willing to countenance sufficiently high values of ε. So HL rules can make sense without making any revisions to $P(p)$ or $P(q)$. The question is, Do people revise down $P(p)$ or revise up ε? Oaksford and Wakefield's (2003) data addressed this issue (see the Probabilities and Natural Sampling section). In the indirect estimate for the HL rule, participants provided very high values of ε of, on average, .84. This clearly suggests that our earlier proposal of the revision strategy was an unnecessary carryover from when we did not allow for exceptions in the original dependence model, as was noted by Evans and Over (1996a, 1996b).

Mentioning the Relevance of Probabilities

Oberauer et al. (1999) took issue with the fact that in Oaksford et al.'s (1997) experiments (see the Reduced Array Selection Task section), participants were cued to the relevance of frequency information for card choice. In Oaksford et al.'s (1999) Experiments 1 and 2, one half of the participants were cued to the relevance of probability information by being asked to assess $P(p)$ and $P(q)$ prior to the selection task. The remaining participants were not cued, because they provided this information only after they had performed the selection task. In Oaksford et al. (1999), we presented no analyses of this order manipulation, because there were no effects that indicated that it was a confounding factor. To test Oberauer et al.'s hypothesis, we looked at each significant effect of the probability manipulations in each subgroup of participants.

More participants selected the *not-q* card when $P(p)$ was high than when it was low. This was significant for the participants who performed the probability check before [$\chi^2(1, N = 64) = 4.22$, $p < .025$; all tests were one-tailed] and after [$\chi^2(1, N = 64) = 4.00$, $p < .025$] the selection task. More participants selected the *not-q* card when $P(q)$ was high than when it was low. This difference was *not* significant for the participants who performed the probability check before the selection task [$\chi^2(1, N = 64) = 0.67$, n.s.], but it was significant for those who performed it afterward [$\chi^2(1, N = 64) = 8.33$, $p < .005$]. More participants selected the q card when $P(q)$ was low than when it was high (this was significant only for the comparison between the LL and the LH rules). This difference was *not* significant for the participants who performed the probability check before the selection task [$\chi^2(1, N = 32) = 1.17$, $p = .14$], but it was significant for those who performed it afterward [$\chi^2(1, N = 32) = 3.14$, $p < .025$]. In summary, contrary to Oberauer et al.'s (1999) suggestion, if anything, cuing participants to the relevance of probability manipulations would appear to suppress, rather than facilitate, their effects on people's card selections.

Sample From a Larger Population

A further criticism of the model is that it assumes that people regard the four cards as a sample from a larger population (Laming, 1996; Oberauer et al., 1999; Peter Wason, personal communication, March 1995). However, it is only on this assumption that considerations from the philosophy of science (e.g., Popper, 1959) bear on the task: No meaningful scientific hypothesis has ever been stated over a domain of only four objects. In the summary of early work in this area (Wason & Johnson-Laird, 1972) the selection task was introduced as bearing on "how, psychologically, science is possible," and it was concluded that "one contributory cause must be a preeminent ability to generalise and to test generalisations" (p. 172). Despite the immediate inference that participants were originally intended to view the cards as a sample from a larger population, they may not. If they do not, then whether they were seeking falsificatory or confirmatory evidence, logically they should turn the *p* and the *not-q* cards. Since participants conspicuously refrain from this selection of cards, it seems reasonable to assume that they do not spontaneously interpret the rule in this way. Moreover, as Oaksford et al. (1999) pointed out, when Legrenzi (1971) presented participants with the four cards so that they could look at both sides, only 1 participant out of 30 described the situation by using a conditional. That is, with such a limited domain, the conditional is not the most natural description of the situation.

However, some authors still have insisted that people *should* interpret the rule as applying only to the four cards. Oberauer et al. (1999), for example, argued that conditionals can be both general and specific—for example, *if it rains tomorrow, the game will be canceled* is an example of a specific conditional claim. However, to construct a selection task with these rules, the cards can no longer be interpreted as instances—that is, as individual people or objects. Rather, they must be interpreted as possible states of affairs concerning what happens tomorrow. Moreover, just like the corresponding

counterfactual, *if it had rained yesterday, the game would have been canceled*, whether one should believe such a rule seems to depend on the existence of a law or social convention that games of this kind do not take place if it rains (Goodman, 1954). Consequently, whether one should believe it or not depends on the evidence one has for the generalization, *if it rains then games of this kind are canceled.* And this need not involve looking at data but may simply involve remembering the laws of the game—for example, if the game in question is tennis, one will be strongly inclined to believe the claim. In sum, the mere fact that conditionals can be used to make specific claims does not mean it would be at all natural to interpret the rules in a selection task as specific. Moreover, even if they were so interpreted, whether one believes specific conditionals of this form may rely on the truth or falsity of a related generalization that may be assumed to be the real rule under test.

Individual Differences

Green (2000) and Evans (1999) both questioned whether the probabilistic approach is sufficient as an explanation of human reasoning performance. Both commentators argued that there is clear evidence of deductive competence. For example, Green (2000) noted that although people are sensitive to the believability of conclusions (see also Oaksford et al., 2000), these effects are far stronger on invalid than on valid conclusions. Moreover, Green (1995a, 1995b; Green & Larking, 1995) has shown that some participants do construe the selection task logically, and Stanovich and West (1998) have shown that a subgroup (around 10%) of participants with high intelligence are capable of logical performance. That is, it appears that people have some sensitivity to the notion of logical validity.

However, as Stanovich and West's (1998) results show, this may be as few as 10% of students at a top rank university (their research was conducted at the University of Toronto). That is, assuming that only the top 1% of the population ever attends such institutions (and even this is probably too liberal an estimate), logical performance may be seen in only as few as 0.1% of the population. Moreover, even the behavior of this elite band does not necessarily implicate an underlying innate logical competence. It could simply reflect an accumulation of experience showing that these particular inferences seem to work in the real world more often than do others. Or it could reflect a learned ability acquired at school while learning, for example, mathematics or IT, which may account for its association with IQ as measured by the SAT (Stanovich & West, 1998).

These results bring us back to one of the issues with which we began: the balance of System 1 and System 2 processes in reasoning. The results of our model comparison appeared to showed that the System 1 information gain model can explain more of the data than can System 2 logic-based models. The latter may be needed only to explain the performance of this elite band of very high IQ participants. Of course, even here these psychologic accounts are partly redundant, because they are designed largely to explain errors in logical reasoning, not successes. Logical success can be explained simply as acquired logical knowledge, which need imply little for the fundamental organization of our cognitive architecture.

CONCLUSION

In this paper, we have presented a revised version of the information gain model and have shown that it provides good fits to the data originally modeled by Oaksford and Chater (1994). We then reviewed the recent literature on the selection task and showed that the revised model can provide good fits to the much of the data and, moreover, addresses a wide range of theoretical criticisms of the model that have been suggested since it first appeared. We have also addressed a variety of findings that have been claimed to be inconsistent with the model, and we have argued in each case that the results do not discriminate against the information gain account to anything like the degree claimed. We believe that the intense scrutiny that the information gain model has undergone in the literature since it first appeared has been extremely healthy. These criticisms not only have resulted in a more coherent model, but also have led to the development of other testable models of the data selection behavior observed in Wason's (1966) selection task. Even if these models are not the last word on that task, echoing Green and Over (2000), the experimental results we have reviewed also reveal that any account of data selection is going to have to explain probabilistic effects such as those predicted by these models.

More generally, the success of these models has some important consequences. First, it is important to realize that there may be more than one normative model of task performance and that deciding which is the most appropriate is an empirical issue—that is, computational-level theories must be normatively correct and descriptively adequate. Second, the information gain model and its variants demonstrate that the behavior observed on the selection task is rational; there is no need to invoke performance errors of the type appealed to by logical approaches to explain the data. Third, these analyses show that the Wason selection task is better viewed as an inductive task, which was how it was first introduced into the literature. Fourth, the mounting evidence on the importance of the rarity assumption shows that people's hypothesis-testing behavior is well adapted to the environment. Finally, the model fits seem to show that high-level System 2 processes may not be required to explain these selection task results; they can all be explained by a computational-level theory of lower level System 1 processes. Following Lloyd Morgan's (1894) canon, if some function can be explained at the lower level, that is probably the level at which it should be explained, since there seems little point attributing people with more high-level cognitive equipment than is needed to explain their behavior.

REFERENCES

Adams, E. (1966). Probability and the logic of conditionals. In J. Hintikka & P. Suppes (Eds.), *Aspects of inductive logic* (pp. 265-316). Amsterdam: North-Holland.

Adams, E. (1975). *The logic of conditionals: An application of probability theory to deductive logic*. Dordrecht: Reidel.

Almor, A., & Sloman, S. A. (1996). Is deontic reasoning special? *Psychological Review*, **103**, 374-380.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory & Cognition*, **23**, 510-524.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, **6**, 57-86.

Beattie, J., & Baron, J. (1988). Confirmation and matching biases in hypothesis testing. *Quarterly Journal of Experimental Psychology*, **40A**, 269-297.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analyses*. New York: Springer-Verlag.

Braine, M. D. S., & O'Brien, D. P. (1998). *Mental logic*. London: Erlbaum.

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition*, **31**, 1-21.

Chater, N. (1995). Neural networks: The new statistical models of mind. In J. P. Levy, D. Bairaktaris, J. A. Bullinaria, & P. Cairns (Eds.), *Connectionist models of memory and language* (pp. 207-227). London: UCL Press.

Chater, N., Crocker, M., & Pickering, M. (1998). The rational analysis of inquiry: The case of parsing. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 441-468). Oxford: Oxford University Press.

Chater, N., & Oaksford, M. (1990). Autonomy, implementation and cognitive architecture: A reply to Fodor and Pylyshyn. *Cognition*, **34**, 93-107.

Chater, N., & Oaksford, M. (1999a). Information gain vs. decision-theoretic approaches to data selection: Response to Klauer (1999). *Psychological Review*, **106**, 223-227.

Chater, N., & Oaksford, M. (1999b). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, **38**, 191-258.

Chater, N., & Oaksford, M. (2001). Human rationality and the psychology of reasoning: Where do we go from here? *British Journal of Psychology*, **92**, 193-216.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality & Social Psychology*, **58**, 545-567.

Chrostowski, J. J., & Griggs, R. A. (1985). The effects of problem content, instructions and verbalisation procedure on Wason's selection task. *Current Psychological Research & Reviews*, **4**, 99-107.

Clocksin, W. F., & Mellish, C. S. (1984). *Programming in Prolog*. Berlin: Springer-Verlag.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral & Brain Sciences*, **4**, 317-370.

Comrie, B. (1986). Conditionals: A typology. In E. C. Traugott, A. ter Meulen, J. S. Reilly, & C. A. Ferguson (Eds.), *On conditionals* (pp. 77-99). Cambridge: Cambridge University Press.

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.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition*, **19**, 274-282.

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little, Brown.

Earman, J. (1992). *Bayes or bust?* Cambridge, MA: MIT Press.

Evans, J. St. B. T. (1977). Toward a statistical theory of reasoning. *Quarterly Journal of Experimental Psychology*, **29**, 621-635.

Evans, J. St. B. T. (1983). Linguistic determinants of bias in conditional reasoning. *Quarterly Journal of Experimental Psychology*, **35A**, 635-644.

Evans, J. St. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, **75**, 451-468.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Erlbaum.

Evans, J. St. B. T. (1999). Rational analysis of illogical reasoning [Review of *Rationality in a uncertain world: Essays on the cognitive science of human reasoning*]. *Contemporary Psychology*, **44**, 461-463.

Evans, J. St. B. T., Clibbens, J., & Rood, B. (1996). The role of implicit and explicit negations in conditional reasoning bias. *Journal of Memory & Language*, **35**, 392-409.

Evans, J. St. B. T., & Handley, S. J. (1999). The role of negation in conditional inference. *Quarterly Journal of Experimental Psychology*, **52A**, 739-770.

Evans, J. St. B. T., & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology*, **64**, 391-397.

Evans, J. St. B. T., & Over, D. E. (1996a). *Rationality and reasoning*. Hove, U.K.: Psychology Press.

Evans, J. St. B. T., & Over, D. E. (1996b). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, **103**, 356-363.

Fedorov, V. V. (1972). *Theory of optimal experiments*. London: Academic Press.

Feeney, A., & Handley, S. J. (2000). The suppression of *q* card selections: Evidence for deductive inference in Wason's selection task. *Quarterly Journal of Experimental Psychology*, **53A**, 1224-1242.

Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition*, **77**, 1-79.

Fiedler, K., & Hertel, G. (1994). Content-related schemata versus verbal-framing effects in deductive reasoning. *Social Cognition*, **12**, 129-147.

Gebauer, G., & Laming, D. (1997). Rational choices in Wason's selection task. *Psychological Research*, **60**, 284-293.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, **102**, 684-704.

Girotto, V., Mazzocco, A., & Cherubini, P. (1992). Judgements of deontic relevance in reasoning: A reply to Jackson and Griggs. *Quarterly Journal of Experimental Psychology*, **45**, 547-575.

Gluck, M. A., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory & Language*, **27**, 166-195.

Goodman, N. (1954). *Fact, fiction and forecast*. Cambridge, MA: Harvard University Press.

Green, D. W. (1995a). The abstract selection task: Thesis, antithesis and synthesis. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning: Essays in honour of Peter Wason* (pp. 171-186). Hove, U.K.: Erlbaum.

Green, D. W. (1995b). Externalisation, counter-examples and the abstract selection task. *Quarterly Journal of Experimental Psychology*, **48A**, 424-446.

Green, D. W. (2000). [Review of *Rationality in a uncertain world: Essays on the cognitive science of human reasoning*]. *Quarterly Journal of Experimental Psychology*, **53A**, 281-283.

Green, D. W., & Larking, R. (1995). The locus of facilitation in the abstract selection task. *Thinking & Reasoning*, **1**, 183-199.

Green, D. W., & Over, D. E. (1997). Causal inference, contingency tables and the selection task. *Current Psychology of Cognition*, **16**, 459-487.

Green, D. W., & Over, D. E. (1998). Reaching a decision: A reply to Oaksford. *Thinking & Reasoning*, **4**, 231-248.

Green, D. W., & Over, D. E. (2000). Decision theoretical effects in testing a causal conditional. *Current Psychology of Cognition*, **19**, 51-68.

Green, D. W., Over, D. E., & Pyne, R. A. (1997). Probability and choice in the selection task. *Thinking & Reasoning*, **3**, 209-236.

Griggs, R. A. (1984). Memory cueing and instructional effects on Wason's selection task. *Current Psychological Research & Reviews*, **3**, 3-10.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, **73**, 407-420.

Hardman, D. (1998). Does reasoning occur on the selection task: A

comparison of relevance-based theories. *Thinking & Reasoning*, **4**, 353-376.

Hattori, M. (1999). The effects of probabilistic information in Wason's selection task: An analysis of strategy based on the ODS model. *Proceedings of the 16th Annual Meeting of the Japanese Cognitive Science Society*, **16**, 623-626.

Hattori, M. (2002). A quantitative model of optimal data selection in Wason's selection task. *Quarterly Journal of Experimental Psychology*, **55A**, 1241-1272.

Hoch, S. J., & Tschirgi, J. E. (1985). Logical knowledge and cue redundancy in deductive reasoning. *Memory & Cognition*, **13**, 453-462.

Horwich, P. (1982). *Probability and evidence*. Cambridge: Cambridge University Press.

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

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

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

Johnson-Laird, P. N., & Wason, P. C. (1970). Insight into a logical relation. *Quarterly Journal of Experimental Psychology*, **22**, 49-61.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, **47**, 263-291.

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

Klauer, K. C. (1999). On the normative justification for information gain in Wason's selection task. *Psychological Review*, **106**, 215-222.

Krauth, J. (1982). Formulation and experimental verification of models in propositional reasoning. *Quarterly Journal of Experimental Psychology*, **34A**, 285-298.

Laming, D. (1996). On the analysis of irrational data selection: A critique of Oaksford and Chater (1994). *Psychological Review*, **103**, 364-373.

Legrenzi, P. (1971). Discovery as a means to understanding. *Quarterly Journal of Experimental Psychology*, **23**, 417-422.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, **27**, 986-1005.

Loehle, C. (2000). *Global Optimization 4.0* [Computer program]. Naperville, IL: Loehle Enterprises.

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

Mackie, J. L. (1963). The paradox of confirmation. *British Journal for the Philosophy of Science*, **38**, 265-277.

Manktelow, K. I., & Evans, J. St. B. T. (1979). Facilitation of reasoning by realism: Effect or non-effect. *British Journal of Psychology*, **70**, 477-488.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

McClelland, J. L. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 21-53). Oxford: Oxford University Press.

McKenzie, C. R. M. (2000). *Examining the rarity assumption and its implications*. Unpublished manuscript, La Jolla, CA: University of California, San Diego, Department of Psychology.

McKenzie, C. R. M., Ferreira, V. S., Mikkelsen, L. A., McDermott, K. J., & Skrable, R. P. (2001). Do conditional statements target rare events? *Organizational Behavior & Human Decision Processes*, **85**, 291-309.

McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin & Review*, **7**, 360-366.

Miller, D. (1994). *Critical rationalism: A restatement and a defence*. Chicago, IL: Open Court.

Morgan, C. L. (1894). *An introduction to comparative psychology*. London: Scott.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning*, **4**, 231-248.

Myerson, J., & Miezen, F. M. (1980). The kinetics of choice: An operant systems analysis. *Psychological Review*, **87**, 160-174.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79-95.

Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking & Reasoning*, **2**, 1-32.

Oaksford, M. (1989). *Cognition and inquiry: The pragmatics of conditional reasoning*. Unpublished doctoral thesis, University of Edinburgh, Centre for Cognitive Science.

Oaksford, M. (1998). Task demands and revising probabilities in the selection task. *Thinking & Reasoning*, **4**, 179-186.

Oaksford, M., & Chater, N. (1991). Against logicist cognitive science. *Mind & Language*, **6**, 1-38.

Oaksford, M., & Chater, N. (1993). Reasoning theories and bounded rationality. In K. I. Manktelow & D. E. Over (Eds.), *Rationality* (pp. 31-60). London: Routledge.

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 & Reasoning*, **1**, 121-152.

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review*, **103**, 381-391.

Oaksford, M., & Chater, N. (1998a). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning*. Hove, U.K.: Psychology Press.

Oaksford, M., & Chater, N. (1998b). A revised rational analysis of the selection task: Exceptions and sequential sampling. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 372-398). Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, **5**, 349-357.

Oaksford, M., Chater, N., & Grainger, B. (1999). Probabilistic effects in data selection. *Thinking & Reasoning*, **5**, 193-243.

Oaksford, M., Chater, N., Grainger, B., & Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 441-458.

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 883-899.

Oaksford, M., & Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 835-854.

Oaksford, M., & Wakefield, M. (2003). Data selection and natural sampling: Probabilities do matter. *Memory & Cognition*, **31**, 143-154.

Oberauer, K., Wilhelm, O., & Rosas Diaz, R. (1999). Bayesian rationality for the Wason selection task? A test of optimal data selection theory. *Thinking & Reasoning*, **5**, 115-144.

Osman, M., & Laming, D. (2001). Misinterpretation of conditional statements in Wason's selection task. *Psychological Research*, **65**, 128-144.

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

Over, D. E., & Jessop, A. (1998). Rational analysis of causal conditionals and the selection task. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 399-414). Oxford: Oxford University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufman.

Pearl, J. (2000). *Causality*. Cambridge: Cambridge University Press.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, **106**, 643-675.

Pollard, P., & Evans, J. St. B. T. (1981). The effect of prior belief in reasoning: An associationist interpretation. *British Journal of Psychology*, **72**, 73-82.

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*. London: Hutchinson.

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. Berlin: Springer-Verlag.

Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.
Sakamoto, Y., & Aikake, H. (1978). Analysis of cross-classified data by AIC. *Annals of the Institute of Statistical Mathematics: Pt. B*, **30**, 185-197.
Schroyens, W., Schaeken, W., Fias, W., & d'Ydewalle, G. (2000). Heuristic and analytic processes in conditional reasoning with negatives. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 1713-1734.
Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General*, **110**, 101-120.
Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.
Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, **57**, 31-95.
Stanovich, K. E., & West, R. F. (1998). Cognitive ability and variation in selection task performance. *Thinking & Reasoning*, **4**, 193-230.
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral & Brain Sciences*, **23**, 645-726.
Stein, E. (1996). *Without good reason*. Oxford: Oxford University Press.
Stich, S. (1985). Could man be an irrational animal? *Synthèse*, **64**, 115-135.
Stich, S. (1990). *The fragmentation of reason*. Cambridge, MA: MIT Press.
Valentine, E. R. (1985). The effect of instructions on performance in the Wason selection task. *Current Psychological Research & Reviews*, **4**, 214-223.
Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology* (pp. 135-151). Harmondsworth, U.K.: Penguin.
Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, **20**, 273-281.
Wason, P. C., & Green, D. W. (1984). Reasoning and mental representation. *Quarterly Journal of Experimental Psychology*, **36A**, 597-610.
Wason, P. C., & Johnson-Laird, P. N. (1970). A conflict between selecting and evaluating information in an inferential task. *British Journal of Psychology*, **61**, 509-515.
Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.
Wiener, N. (1948). *Cybernetics*. New York: Wiley.
Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Newbury Park, CA: Sage.
Wolfram, S. (1999). *Mathematica 4.0* [Computer program]. Cambridge: Cambridge University Press.
Yachanin, S. A. (1986). Facilitation in Wason's selection task. *Current Psychological Research & Reviews*, **5**, 20-29.

NOTES

1. Krauth (1982) extended the stochastic model of Evans (1977) to explain negations paradigm data. He also provided detailed model fits. However, these models parameterized various tendencies to verification, falsification, and matching or degree of insight into the task. As psychological constructs, these notions are now outdated. In mental logic and mental models theory, these notions have been replaced by the proposal that people adopt different interpretations of the task rule, as we discuss in the text. Consequently, although these models may provide good fits to the data, theoretically they are no longer interesting. In particular, the parameters refer to probabilities of, for example, *being in a verifying state*, but these are not parameters that we can vary experimentally, unlike the parameters of the information gain model.

2. Are the disinterested and the decision-theoretic approaches related? One might imagine that when the costs of making different errors are equal (or more specifically, both 0), the decision-theoretic approach collapses into the disinterested approach. Under these circumstances (particularly when these costs are both 0; see Klauer, 1999, Appendix B), any decision-theoretic approach may certainly be described as *disinterested*, although one might argue that the cost of experimentation still needs to be taken into account in computing a risk function (Klauer, 1999). However, this is also the case for *disinterested* approaches—that is, it explains why people do not turn all the cards with some positive information value (Chater & Oaksford, 1999b). However, it is important to note that, under the conditions mentioned, existing decision-theoretic approaches do not reduce formally to the existing disinterested approaches. This is because they use different notions of informativeness (as Klauer, 1999, observed; even though Oaksford & Chater, 1996, showed that expected information gain is equivalent to expected Kullback–Leibler information, the Kullback–Leibler numbers used in the optimal Bayes's procedure are different again). Moreover, even if the costs of making different errors were equal (but not both 0), they would diverge in their predictions when people do not believe the rule (see the main text).

3. Could we have captured degrees of dependency other than by introducing an exceptions parameter? McKenzie and Mikkelsen (2000) have suggested that the phi coefficient (Siegel & Castellan, 1988) could be used. This statistic provides a measure of the correlation between two dichotomous variables in a 2×2 contingency table. One argument for not using the phi coefficient is as follows. Looking at Table 1, if the p, *not-q* cell is 0 and the other cells are equiprobable, phi = .5, whereas $\varepsilon = 0$. However, if the *not-p*, q cell were 0 and the other cells were equiprobable, again phi = .5, but now $\varepsilon = .5$. That is, the phi coefficient is unable to distinguish between a case in which, for example, turning the key always starts the car (there are no exceptions) and a case in which it is at chance levels whether turning the key starts the car (half of all cases are exceptions). This seems undesirable in a measure of conditional dependency (we thank Masasi Hattori, personal communication, June 2001, for this point).

(*Continued on next page*)

APPENDIX

Table A1
The Fits to the Abstract Selection Task Data in Oaksford and Chater's (1994) Meta-Analysis, Showing the Observed Probability of Selecting Each Card [Obs. *P*(Sel.)] and the Probability of Selecting Each Card Predicted by the Information Gain Model [Pred. *P*(Sel.) IG] and the Mental Logic and Mental Models Approaches [Pred. *P*(Sel.) MM]

		Obs. *P*(Sel.)				Pred. *P*(Sel.) IG							Pred. *P*(Sel.) MM					
Study	Experiment	p	$\overline{p}$	q	$\overline{q}$	p	$\overline{p}$	q	$\overline{q}$	$P(p)$	$P(q)$	$G^2(2)$	$\overline{p}$	q	$\overline{q}$	P_c	P_f	$G^2(2)$
Wason (1968)	1/experimental	1.0	.17	.78	.28	.92	.13	.68	.16	.18	.22	5.8*	.27	.63	.16	.23	.25	4.4*
	1/control	1.0	.06	.69	.12	.95	.09	.76	.09	.04	.05	2.4*	.12	.63	.07	.32	.11	3.9*
	2	1.0	.12	.50	.12	.98	.09	.48	.10	.05	.10	1.1*	.12	.50	.12	.48	.15	6.9*
Evans & Lynch (1973)	Single	.88	.08	.50	.33	.89	.17	.48	.30	.28	.37	1.7*	.17	.40	.26	.53	.29	1.0*
Manktelow & Evans (1979)	1/abstract	.96	.12	.62	.33	.91	.15	.55	.23	.24	.31	2.8*	.23	.49	.23	.40	.29	2.2*
	2/abstract	.83	.17	.67	.25	.83	.18	.66	.23	.28	.31	0.1*	.24	.58	.17	.33	.25	0.6*
	3/abstract	.88	.38	.56	.38	.78	.23	.53	.35	.35	.40	2.6*	.40	.53	.35	.39	.45	1.3*
	4/abstract	.69	.00	.62	.12	.82	.18	.73	.20	.26	.27	9.3	.07	.55	.06	.39	.08	6.7*
	5/abstract	.88	.12	.81	.06	.89	.11	.84	.11	.11	.12	0.6*	.15	.79	.04	.18	.10	0.7*
Griggs & Cox (1982)	1/Trial 1	1.0	.06	.56	.06	.97	.09	.60	.10	.06	.09	3.2*	.07	.56	.06	.44	.12	3.8*
	1/Trial 2	1.0	.06	.69	.06	.96	.09	.72	.09	.02	.03	2.6*	.08	.67	.04	.31	.07	3.9*
	3/Trial 1	.70	.30	.70	.40	.67	.29	.63	.31	.38	.39	1.2*	.39	.56	.30	.30	.41	5.2*
	3/Trial 2	.65	.30	.75	.30	.64	.29	.72	.26	.36	.35	0.3*	.37	.64	.21	.24	.34	8.2*
Griggs (1984)	non-mem./T–F	.84	.16	.52	.08	.90	.15	.61	.21	.23	.28	4.9*	.13	.55	.11	.44	.15	3.3*
	non-mem./vio.	.92	.16	.76	.16	.90	.13	.74	.15	.18	.20	0.4*	.22	.69	.10	.23	.18	0.4*
Chrostowski & Griggs (1985)	non-mem./vio.	.87	.12	.77	.15	.88	.14	.77	.15	.19	.20	0.2*	.18	.69	.08	.23	.15	0.3*
	non-mem./T–F	.97	.07	.78	.08	.95	.09	.76	.09	.04	.05	2.3*	.11	.74	.04	.22	.08	5.0*
Hoch & Tschirgi (1985)*	bachelor's	.88	.24	.60	.40	.83	.21	.53	.32	.32	.38	1.9*	.31	.49	.33	.40	.40	0.02*
Valentine (1985)	AA	.83	.12	.58	.25	.86	.18	.60	.25	.28	.33	0.6*	.19	.50	.19	.42	.24	0.7*
Yachanin (1986)	2/widgit/vio.	1.0	.15	.80	.20	.94	.11	.73	.12	.12	.14	4.5*	.23	.69	.10	.20	.19	4.7*
	2/widgit/test	.95	.20	.70	.30	.89	.16	.63	.22	.25	.29	2.5*	.29	.58	.21	.30	.29	0.8*
Beattie & Baron (1988)	1/4-card, +ve	.94	.06	.56	.06	.98	.09	.59	.09	.02	.04	1.4*	.06	.55	.05	.44	.07	0.9*
	2/4-card, +ve	1.0	.06	.62	.50	.92	.15	.45	.29	.26	.37	9.1*	.24	.40	.35	.46	.39	7.8*
	3/4-card, +ve	1.0	.00	.62	.12	.96	.11	.64	.13	.12	.16	4.4*	.07	.55	.06	.39	.08	5.8*
Cosmides (1989)	1 & 2	.96	.21	.31	.44	.93	.15	.30	.42	.28	.46	2.3*	.20	.32	.44	.65	.47	4.2*
	3 & 4	.96	.23	.48	.52	.88	.18	.37	.42	.31	.44	9.0*	.30	.39	.47	.53	.51	3.1*
Girotto et al. (1992)	1/arbitrary rule	.83	.17	.58	.25	.85	.18	.60	.26	.29	.34	0.1*	.22	.52	.20	.41	.26	0.7*
	2/arbitrary rule	.96	.29	.54	.33	.88	.17	.50	.30	.29	.37	4.1*	.32	.51	.31	.42	.39	2.5*
	3/arbitrary rule	.79	.29	.42	.54	.75	.25	.38	.51	.39	.48	0.6*	.32	.38	.52	.55	.56	2.2*
	4/arbitrary rule	.80	.10	.60	.25	.84	.18	.62	.25	.28	.33	1.4*	.18	.51	.17	.41	.22	1.6*
Oaksford & Stenning (1992)	2/abstract	.79	.25	.62	.21	.79	.21	.65	.26	.31	.34	0.6*	.27	.60	.18	.34	.27	3.2*
	3/colored shape	.62	.29	.62	.29	.64	.31	.64	.31	.39	.39	0.1*	.33	.57	.25	.35	.35	12.2
	3/vowel-even	.96	.17	.71	.29	.90	.15	.63	.20	.23	.27	3.0*	.26	.58	.19	.30	.27	1.7*
	3/control	.88	.17	.50	.17	.91	.15	.55	.24	.25	.31	1.2*	.17	.50	.17	.47	.21	1.1*

Note—The best-fitting parameter values and the values of the log-likelihood ratio (G^2) are also shown. For the information gain model, $P(M_I) = .5$ and $\varepsilon = .1$ for all model fits. For all the mental logic/mental models fits, the predicted probability of selecting the p card is not shown because it is always .888. For all studies using Evans's (Evans & Lynch, 1973) negations paradigm, only the data for the affirmative rule are included. Studies were included only where individual card selection frequencies were reported or could be inferred from exhaustive reporting of card combinations. $\overline{p}$, *not-p* card; $\overline{q}$, *not-q* card; MM, mental logic and mental models theories; non-mem., no memory cuing; T, true; F, false; vio., violation condition; AA, affirmative antecedent and affirmative consequent condition; +ve, affirmative consequent condition. *Model cannot be rejected at the .01 level of significance.

Table A2
The Fits of the Information Gain Model to the Negations Paradigm Data in Oaksford and Chater's (1994) Meta-Analysis, Showing the Observed Probability of Selecting Each Card [Obs. *P*(Sel.)] and the Probability of Selecting Each Card Predicted by the Information Gain Model [Pred. *P*(Sel.) IG]

		Obs. *P*(Sel.)				Pred. *P*(Sel.) IG						
Study	Rule	p	$\overline{p}$	q	$\overline{q}$	p	$\overline{p}$	q	$\overline{q}$	*P*(TA)	*P*(TC)	$G^2(2)$
Evans & Lynch (1973)	*If p then q*	.88	.08	.50	.33	.89	.17	.48	.30	.28	.37	1.7*
	If p then not-q	.92	.04	.08	.58	.96	.10	.12	.63	.18	.68	2.7*
	If not-p then q	.54	.21	.71	.08	.60	.28	.82	.20	.33	.30	5.2*
	If not-p then not-q	.83	.13	.33	.25	.90	.16	.42	.34	.28	.40	2.9*
Manktelow & Evans (1979), Experiment 1	*If p then q*	.96	.13	.63	.33	.91	.15	.55	.23	.24	.31	2.8*
	If p then not-q	1.0	.08	.21	.75	.95	.10	.12	.69	.19	.71	4.6*
	If not-p then q	.58	.29	.58	.42	.62	.34	.56	.38	.42	.43	0.6*
	If not-p then not-q	.54	.46	.29	.75	.49	.41	.28	.72	.52	.60	0.6*
Manktelow & Evans (1979), Experiment 2	*If p then q*	.83	.17	.67	.25	.83	.18	.66	.23	.28	.31	0.1*
	If p then not-q	.96	.04	.33	.75	.91	.13	.17	.66	.28	.60	7.7*
	If not-p then q	.79	.29	.71	.50	.72	.27	.57	.34	.37	.40	5.1*
	If not-p then not-q	.83	.21	.54	.67	.78	.24	.37	.51	.38	.48	6.0*
Oaksford & Stenning (1992), Unpublished Control 1	*If p then q*	.79	.25	.63	.21	.79	.21	.65	.26	.31	.34	0.6*
	If p then not-q	.83	.25	.17	.67	.82	.19	.22	.69	.36	.58	1.1*
	If not-p then q	.83	.29	.79	.29	.76	.21	.72	.23	.30	.31	2.7*
	If not-p then not-q	.71	.33	.38	.58	.68	.30	.36	.57	.43	.51	0.3*
Oaksford & Stenning (1992), Experiment 3, Control	*If p then q*	.75	.21	.54	.29	.78	.23	.57	.32	.34	.38	0.3*
	If p then not-q	.75	.13	.21	.46	.83	.20	.30	.55	.35	.51	3.9*
	If not-p then q	.58	.33	.71	.42	.58	.36	.64	.32	.41	.40	1.5*
	If not-p then not-q	.88	.25	.54	.42	.82	.21	.49	.36	.33	.40	1.3*
Oaksford & Stenning (1992), Unpublished Control 2	*If p then q*	.71	.17	.50	.17	.79	.22	.61	.29	.33	.36	4.3*
	If p then not-q	.79	.13	.17	.50	.86	.18	.26	.58	.34	.53	3.2*
	If not-p then q	.67	.21	.46	.46	.71	.28	.45	.45	.40	.46	0.9*
	If not-p then not-q	.83	.13	.63	.42	.83	.20	.55	.31	.32	.37	2.5*

Note—The best-fitting parameter values and the values of the log-likelihood ratio (G^2) are also shown. For the information gain model, $P(M_1) = .5$ and $\varepsilon = .1$ for all model fits. $\overline{p}$, *not-p* card; $\overline{q}$, *not-q* card. *Model cannot be rejected at the .01 level of significance.

Table A3
The Fits to the Negations Paradigm Data in Oaksford and Chater's (1994) Meta-Analysis, Showing the Probability of Selecting Each Card Predicted by the Three-Parameter Information Gain Model [Pred. *P*(Sel.) IG] and the Three-Parameter Mental Logic and Mental Models Approaches [Pred. *P*(Sel.) MM]

		Pred. *P*(Sel.) MM				MM		Pred. *P*(Sel.) IG				IG	
Study	Rule	p	$\overline{p}$	q	$\overline{q}$	Parameter	$G^2(13)$	p	$\overline{p}$	q	$\overline{q}$	Parameter	$G^2(13)$
Evans & Lynch (1973)	*If p then q*	.86	.04	.51	.24	$P_c = .49$	31.6	.93	.14	.47	.26	$P(p) = .24$	26.3*
	If p then not-q	.86	.04	.33	.42	$P_f = .42$		.95	.13	.26	.39	$P(q) = .34$	
	If not-p then q	.68	.22	.51	.24	$P_n = .18$		.60	.30	.76	.24	$P(n) = .12$	
	If not-p then not-q	.68	.22	.33	.42			.81	.22	.39	.46		
Manktelow & Evans (1979), Experiment 1	*If p then q*	.86	.10	.58	.40	$P_c = .42$	13.6*	.91	.16	.35	.38	$P(p) = .28$	22.5*
	If p then not-q	.86	.10	.26	.72	$P_f = .72$		.91	.14	.18	.64	$P(q) = .43$	
	If not-p then q	.54	.41	.58	.40	$P_n = .32$		.54	.40	.61	.35	$P(n) = .16$	
	If not-p then not-q	.54	.41	.26	.72			.66	.28	.25	.73		
Manktelow & Evans (1979), Experiment 2	*If p then q*	.86	.17	.59	.40	$P_c = .41$	24.3*	.88	.18	.44	.34	$P(p) = .30$	30.1
	If p then not-q	.86	.17	.43	.56	$P_f = .56$		.90	.16	.29	.48	$P(q) = .40$	
	If not-p then q	.70	.33	.59	.40	$P_n = .16$		.67	.30	.60	.33	$P(n) = .09$	
	If not-p then not-q	.70	.33	.43	.56			.77	.25	.37	.51		
Oaksford & Stenning (1992), Unpublished Control 1	*If p then q*	.86	.16	.62	.37	$P_c = .38$	23.4*	.84	.20	.48	.35	$P(p) = .32$	21.9*
	If p then not-q	.86	.16	.43	.57	$P_f = .57$		.87	.18	.31	.50	$P(q) = .40$	
	If not-p then q	.66	.36	.62	.37	$P_n = .20$		.59	.35	.64	.32	$P(n) = .09$	
	If not-p then not-q	.66	.36	.43	.57			.71	.28	.40	.51		
Oaksford & Stenning (1992), Experiment 3, Control	*If p then q*	.86	.14	.58	.34	$P_c = .42$	20.4*	.81	.22	.53	.33	$P(p) = .33$	12.3*
	If p then not-q	.86	.14	.45	.48	$P_f = .48$		.85	.19	.37	.44	$P(q) = .39$	
	If not-p then q	.72	.28	.58	.34	$P_n = .14$		.62	.33	.66	.31	$P(n) = .06$	
	If not-p then not-q	.72	.28	.45	.48			.72	.28	.46	.44		
Oaksford & Stenning (1992), Unpublished Control 2	*If p then q*	.86	.12	.48	.34	$P_c = .52$	24.5*	.83	.21	.47	.37	$P(p) = .33$	23.0*
	If p then not-q	.86	.12	.39	.43	$P_f = .43$		.85	.20	.38	.43	$P(q) = .41$	
	If not-p then q	.77	.21	.48	.34	$P_n = .09$		.74	.26	.54	.36	$P(n) = .04$	
	If not-p then not-q	.77	.21	.39	.43			.78	.24	.44	.44		

Note—The best-fitting parameter values and the values of the log-likelihood ratio (G^2) are also shown. For the information gain model, $P(M_I) = .5$ and $\varepsilon = .1$ for all model fits. $\overline{p}$, *not-p* card; $\overline{q}$, *not-q* card. *Model cannot be rejected at the .01 level of significance.

(Manuscript received October 24, 2000;
revision accepted for publication May 17, 2002.)