THINKING AND REASONING, 1999, *5* (3), 193–243

Probabilistic Effects in Data Selection

Mike Oaksford
Cardiff University, UK

Nick Chater and Becki Grainger
University of Warwick, UK

Four experiments investigated the effects of probability manipulations on the indicative four card selection task (Wason, 1966, 1968). All looked at the effects of high and low probability antecedents (p) and consequents (q) on participants' data selections when determining the truth or falsity of a conditional rule, if p then q. Experiments 1 and 2 also manipulated believability. In Experiment 1, 128 participants performed the task using rules with varied contents pretested for probability of occurrence. Probabilistic effects were observed which were partly consistent with some probabilistic accounts but not with non-probabilistic approaches to selection task performance. No effects of believability were observed, a finding replicated in Experiment 2 which used 80 participants with standardised and familiar contents. Some effects in this experiment appeared inconsistent with existing probabilistic approaches. To avoid possible effects of content, Experiments 3 (48 participants) and 4 (20 participants) used abstract material. Both experiments revealed probabilistic effects. In the Discussion we examine the compatibility of these results with the various models of selection task performance.

INTRODUCTION

Results in the psychology of reasoning appear to show that people make systematic errors on tasks with apparently straightforward logical solutions. The task most often used to illustrate this point in both the philosophical and the psychological literature is Wason's (1966, 1968) selection task. In the selection task an experimenter presents participants with four cards, each with a number on

Requests for reprints should be sent to Mike Oaksford, School of Psychology, Cardiff University, PO Box 901, Cardiff, CF1 3YG, UK, Wales. Email: oaksford@cardiff.ac.uk.

These experiments were conducted while the first author was at the Department of Psychology, University of Warwick, Coventry, UK, and the second author was at the Department of Experimental Psychology, University of Oxford, Oxford, UK.

We thank Jonathan Evans, David Green, Keith Holyoak, Donald Laming, Ray Nickerson, David Over, and Dan Sperber for their many helpful comments on this paper.

one side and a letter on the other, and a rule of the form *if p then q*, e.g. *if there is a vowel on one side* (*p*), *then there is an even number on the other side* (*q*). This rule is in the indicative mood—it putatively describes how the world is. The four cards show an "A"(*p* card), a "K"(*not-p* card), a "2"(*q* card), and a "7"(*not-q* card). Participants have to select those cards that they must turn over to determine whether the rule is true or false. Logically participants should select only the *p* and the *not-q* cards, i.e. those cards with the potential to reveal a falsifying instance (Popper, 1959). However, as few as 4% of participants make this a response, other responses being far more common: *p* and *q* cards (46%); *p* card only (33%); *p*, *q* and *not-q* cards (7%); *p* and *not-q* cards (4%) (Johnson-Laird & Wason, 1970). This robust and reliable effect has been widely interpreted to cast doubt on human rationality (Cohen, 1981; Manktelow & Over, 1993; Stich, 1985, 1990) which has raised the selection task to the status of a benchmark against which theories of reasoning are often judged. More recently the selection task has been the focus of attempts to unify the areas of deductive reasoning and decision making (e.g. Evans, Over, & Manktelow, 1993).

All theories of human reasoning have attempted to explain selection task results. There are two hierarchically related distinctions between the different theories. First, there are probabilistic theories (e.g. Evans & Over, 1996a, Kirby, 1994; Klauer, 1999; Nickerson, 1996; Oaksford & Chater, 1994, 1995a, 1996) and non-probabilistic theories (e.g. Cheng & Holyoak, 1985; Cosmides, 1989; Evans, 1984, 1989; Johnson-Laird & Byrne, 1991; Rips, 1994; Sperber, Cara, & Girotto, 1995). Second, the different probabilistic theories make different predictions for the selection task. In particular, some predict no influence of prior belief (Nickerson, 1996; Oaksford & Chater, 1994, 1995a, 1996) whereas some predict more falsificatory responses for disbelieved rules (Evans & Over, 1996a, b; Klauer, 1999). The purpose of these experiments is therefore to investigate whether probabilistic effects occur in the standard indicative form of the selection task and to see whether prior beliefs affect people's behaviour.

In the next two sections we outline the non-probabilistic and the probabilistic approaches to the indicative selection task.

NON-PROBABILISTIC THEORIES OF THE SELECTION TASK

Mental Logic and *PSYCOP*

The mental logic approach (e.g. Braine, 1978; Rips, 1983, 1990) assumes that people reason logically using syntactic rules. Most mental logicians have avoided dealing with selection task results. For example, Rips (1990) has argued that mental logic does not apply to this task, because it is not a logical problem but an inductive, probabilistic problem, in line with probabilistic approaches. However, recently Rips (1994) has argued that his PSYCOP model can explain the selection task within the mental logic framework. In PSYCOP, the pattern of

logical "errors" is modelled 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. So for example, 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 card) means that PSYCOP cannot apply *backward* rules that work from conclusion to premises (as in a PROLOG interpreter, see Clocksin & Mellish, 1983). 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, i.e. *if p then q and if q then p*. For Rips, succeeding on this problem requires proposing assumptions about what is on the backs of the cards so that backward rules can also be applied. Notice that according to the mental logic approach people *should* respond logically—they only fail to do so because of the particular algorithms hypothesised to implement logic in the mind.

Mental Models

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 behaviour 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 only represent the named cases, i.e.

[*p*] *q*
...

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. 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

[*p*] [*q*]
…

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

[*p*] *q*

not-p *not-q*

...

people will realise 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.

Like recent mental logic theories (Rips, 1994), mental models theory assumes that people are capable of reasoning logically in this task but that the psychological mechanisms involved make it difficult to derive the full logical response, i.e. "people are rational [logical] in principle but fallible in practice" (Johnson-Laird & Byrne, 1991, p.19). Both theories also agree that, given certain restrictions, logic provides the computational-level theory (Marr, 1982) of the selection task, i.e. the theory of what people *should* do.

Pragmatic Reasoning Schema and Darwinian Algorithms

Pragmatic reasoning schema theory (Cheng & Holyoak, 1985) was developed to explain the realisation that the "facilitating" effect of thematic material in the selection task was due to the *deontic* nature of the rules used. For example, the rule *if someone is drinking beer, they must be over 21 years of age* (Griggs & Cox, 1983) leads to much higher selections of *p* and *not-q* cards. This happens because this rule provides a prescription for how you are obliged to behave (deontic), rather than a description of how the world is (indicative). A pragmatic reasoning schema contains domain-specific rules that apply when obligations and permissions are described by a rule. Similarly Cosmides (1989) has argued that people possess Darwinian algorithms that enable them to check for *cheaters*, i.e. people who take benefits without paying costs, e.g. under-age drinkers. These theories diverge in their predictions for some rules. However, because they have been applied mainly to the deontic tasks we are not concerned to explicitly test them in this paper.[1] Rather we are concerned that in our subsequent experiments we can discount any facilitation effects being due to the rules involved being interpretable either as obligations or as having a cost–benefit structure.

Relevance and Heuristic Approaches

Two approaches argue that in the selection task people do not reason at all, but instead make relevance judgements of one sort or another. According to Evans (1983, 1984, 1989), relevance judgements are mediated by two linguistic

[1]Although it has been proposed that people may possess many different schemata including one for causal reasoning (Cheng & Holyoak, 1985) we know of no specific proposals for dealing with the probabilistic manipulations we introduce in these experiments.

heuristics (as well as by probabilistic factors, see later). An *if*-heuristic focuses attention on cases that satisfy the antecedent, *p*, biasing people towards selection of the *p* card. A *not*-heuristic focuses attention on items named in the rule, *p* and *q*, biasing people towards selection of the *p* and the *q* cards. This is called the "*not*"-heuristic because it serves to explain people's behaviour when negations are used in the antecedents, *p*, and consequents, *q*, of the rule: they tend to ignore the negations and match the named items (Evans & Lynch, 1973). The combination of these two heuristics can explain the predominant selection patterns in the task.

Recently Sperber, Cara, and Girotto (1995) have argued for a different understanding of relevance using the linguistic theory of Sperber and Wilson (1986). The core of this account is how a *p*, *not-q* response can be made relevant. Conditionals may be relevant for a variety of reasons. One is when the logically equivalent denial—something does *not* exist that is both *p* and *not-q*—is highlighted. This interpretation will lead people to check that such an instance is not on the cards and hence to check the *p* and *not-q* cards. Sperber et al. (1995) argue that this interpretation is highlighted when the *cognitive effects* of adopting it are high and the *cognitive effort* of deriving it is low. They suggest a variety of ways to achieve this. For instance, on the effort side, they suggest making the *not-p* and *not-q* categories easy to represent by allowing the negated categories to be an explicitly introduced positive feature. For example, they may be told that only two letters, A or B, and two numbers, 1 and 2, appear on the cards, so that with respect to the rule if A, then 1, *not-p* = B and *not-q* = 2. This manipulation allows all possible combinations to be readily represented. On the effects side, participants might be told that the *p*, *not-q* instance is diagnostic of a fault, e.g. for a machine printing cards according to the rule if A, then 1, a card with A and 2 on it shows that the machine is not working properly.

PROBABILISTIC THEORIES OF THE SELECTION TASK

A variety of probabilistic approaches to the selection task have recently been proposed (Chater & Oaksford, 1999a; Evans & Over, 1996a, b; Kirby, 1994; Klauer, 1999; Nickerson, 1996; Oaksford & Chater, 1994, 1995a, 1996, 1998a; Over & Evans, 1994, Over & Jessop, 1998). This development is not in itself novel. Many researchers have proposed a similar approach in the past (e.g. Dorling, 1983; Klayman & Ha, 1987; Rips, 1990). What is novel is that specific formal models have been developed which show that this approach can account for a broad range of results on the selection task.

All these probabilistic accounts attempt to explain putative errors by adopting different accounts of what people *should* do in the task. These accounts are based on probability theory rather than on logic. By switching to a different account of what people should do, these approaches argue for a different explanation of the

mismatch between human reasoning performance and logical expectations. Rather than suggesting that people are trying but failing to perform logical inferences because of limitations at the algorithmic level, these accounts suggest that people are succeeding at drawing probabilistic inferences.

Kirby (1994)

Kirby developed a signal detection approach to the task in which the signal to be detected is a card that is inconsistent with the rule, i.e. a *p*, *not-q* instance. According to this account people should choose a card when the posterior odds of an inconsistent outcome exceed a simple function of the utilities associated with a Hit, a False Alarm (*FA*), a Correct Rejection (*CR*), or a Miss (see equation 1). In deriving predictions, Kirby assumes that the utilities on the right hand side of equation (1) remain constant.

$$\frac{P(\text{inconsistent outcome present} \mid C)}{P(\text{inconsistent outcome absent} \mid C)} > \frac{U(\text{CR}) - U(FA)}{U(\text{Hit}) - U(\text{Miss})} \qquad (1)$$

As *q* and *not-p* cards (*C*) have 0 probability of yielding an inconsistent outcome, as with logical accounts, Kirby's predicts that participants should never turn these cards and hence interest centres on the *p* and *not-q* cards.[2] Equation (1) predicts that the posterior odds of finding an inconsistent outcome with the *not-q* card will increase if $P(p)$ is larger, and hence that participants should choose the *not-q* card more frequently. Note, however, that equation (1) predicts no changes for the *q* cards.

Optimal Data Selection

The next four accounts we look at (Evans & Over, 1996a, b; Klauer, 1997; Nickerson, 1996; Oaksford & Chater, 1994, 1995a, 1996; Over & Evans, 1994, Over & Jessop, 1998) can all be encompassed within the optimal experimental design approach (Berger, 1985; Fedorov, 1972) and all share a basic underlying structure that we articulate before highlighting where they disagree.

The basic assumptions shared by all these approaches were specified by Oaksford and Chater (1994). All assume that people are attempting to choose between probabilistic models of the world. In practice two models are usually considered, one in which the rule is true (M_D: a *Dependence* model) and a foil hypothesis in which the two sides of the cards are statistically *I*ndependent (M_I). (See the later section on the *Effects of Prior Beliefs* for some proposals for a different selection of models.) People want to know which model truly describes

[2]This is because for this analysis to make sense the goal must be to avoid false alarms and misses, $U(\text{CR}) > U(FA)$ and $U(Hit) > \text{U}(Miss)$.

the world given the possible data that they can select. This depends on three parameters: the probability of the antecedent $P(p)$; the probability of the consequent, $P(q)$; and the prior probability that the rule is believed true in the first place $P(M_D)$. On the further shared assumption that $P(p)$ and $P(q)$ are low (Oaksford & Chater's, 1994, "rarity" assumption), then the probabilities of what is on the unseen sides of each card follow the distribution shown qualitatively in Table 1. As Oaksford and Chater (1996) argue, the rarity assumption can be readily justified from the literature on Bayesian epistemology (e.g. Horwich, 1982; Howsen & Urbach, 1989). Table 1 shows clearly that under this assumption the *p* and *q* cards are the most informative in discriminating between hypotheses, i.e. these cards show the greatest difference in the probabilities of finding the same outcome under the two models.

The theoretical accounts we look at differ on how to formalise the notion of informativeness. The critical difference concerns whether a "disinterested" or a "decision-theoretic" approach is taken to inquiry (Chater, Crocker, & Pickering, 1998; Chater & Oaksford, 1999a). On the disinterested approach, people are inquiring into the structure of their world with no particular decision problem in mind, their only goal is find out about their world. On this account the only costs that enter into the selection of cards concerns those of turning a card. On the decision-theoretic approach, people are inquiring into their world with a particular decision problem in mind. For example, they might be trying to determine whether they should eat tripe by seeking evidence concerning whether tripe makes you ill. On this account the utilities of various decisions (to eat or not to eat tripe) should clearly affect their data selection activities. The first two probabilistic accounts we look at (Nickerson, 1996; Oaksford & Chater, 1994) take a disinterested approach, whereas the second two (Evans & Over, 1996a; Klauer, 1999) take a decision-theoretic approach.

TABLE 1
Dependence and Independence Models

Turn	*Find*	*Rule true (M_D)*	*Unrelated (M_I)*
p card	*q*	certain	low
	not-q	impossible	high
not-p card	*q*	(very) low	low
	not-q	very high	high
q card	*p*	high	low
	not-p	low	high
not-q card	*p*	impossible	low
	not-p	certain	high

Probabilities of finding a particular outcome on turning each card in the selection task under a model where the rule is true (M_D) and under a model where the sides are unrelated (M_I) and assuming rarity.

Oaksford and Chater (1994). Borrowing from the Bayesian optimal design literature (Fedorov, 1972), Oaksford and Chater (1994) used a measure first introduced by Lindley (1956). They call this measure "expected information gain" because it is a measure of the difference between the information in the prior distribution, $P(M_D)$, and the expected amount of information in the posterior distribution, $P(M_D|D)$, i.e. after selecting some data (a card). Information is defined quantitatively using Shannon-Wiener information (Shannon & Weaver, 1949; Wiener, 1948) and the posteriors are calculated using Bayes' theorem. On the information gain account, selections of the *not-q and* the *q* cards are both sensitive to changes in either $P(p)$ *or* $P(q)$ (Oaksford & Chater, 1994). The main predictions that differentiate the information gain account from Kirby's (1994) are that there should be effects of $P(q)$ (i) on the *q* card such that there are more *q* card selections when $P(q)$ is low, and (ii) on the *not-q* card such that there should be more *not-q* card selections when $P(q)$ is high (see Oaksford & Chater, 1994).

Nickerson (1996). Nickerson (1996) proposed a Bayesian analysis of the selection task that is very similar to Oaksford and Chater (1994) who measure information using the expected value ($E()$) of the difference between the prior ($I(H)$) and posterior information ($I(H|D)$), i.e. information gain (I_g) is defined as follows:

$$I_g = E(I(H|D)–I(H)) \tag{2}$$

Nickerson (1996) instead defines a notion of "impact." Rather than use the prior (P(H)) and posterior ($P(H|D)$) probabilities to compute Shannon-Wiener information, impact is defined as the expected value of the absolute difference between the prior and posterior probabilities, i.e. impact (I_P) is defined as:

$$I_p = E(\,|\,P(H|D)–P(H)\,|\,) \tag{3}$$

Not surprisingly information gain and impact make essentially the same predictions in the selection task.

Evans and Over (1996a). In response to Oaksford and Chater's (1994) article, Evans and Over (1996a) argued that a better measure of the informativeness of a card would be given by the expected absolute value of the log-likelihood ratio:

$$\log\left(\frac{P(D|M_D)}{P(D|M_I)}\right) \tag{4}$$

(4) is the "diagnosticity" of the card, i.e. the extent to which the other side of the card (the data "*D*") discriminates between the two hypotheses. This measure has not been systematically applied to the original selection task, so its predictions have not been fully specified or empirically evaluated.[3] However, Over and Jessop (1998) have used it to model "causal" selection tasks, where they show that it makes identical predictions to information gain (Oaksford & Chater, 1994). However, Evans and Over (1996a) argue informally that there should be strong effects of believability, $P(M_D)$, on selection task performance. We have verified these predictions formally. Specifically, their measure predicts more *not-q* card selections than *q*-card selections when believability is low, even when rarity holds. This prediction contrasts with information gain and impact, which both predict that card selections should be largely independent of believability. As Oaksford and Chater (1996, 1998a) point out, although Evans and Over (1996a; Over & Jessop, 1998) endorse a decision-theoretic approach their measure does not explicitly introduce utilities into the decision to turn cards. However, Klauer (1999) has developed a thoroughgoing decision-theoretic model of the task that captures most of Evans and Over's intuitions.

Klauer (1999). Klauer (1999) argues that information gain is not optimal in one sense discussed by Oaksford and Chater (1996)[4] and he introduces two different optimal decision-theoretic procedures that could apply to the selection task. The measures used in these procedures both embody costs (i) for making different types of epistemic error, i.e. the type I and type II errors of standard statistical hypothesis testing, and (ii) for each experiment (turning each card). These costs combine to create an overall loss function. The decision to turn a card depends on whether the amount of information available exceeds the losses a decision maker is willing to accept. Following work in Bayesian decision theory, Klauer employs Kullback-Liebler information rather than uncertainty reduction to measure information (see equation 5). Equation 5 shows the information provided by experiment *e* (e.g. turning the *p* card) given models (hypotheses) M_i and M_j and different possible outcomes (*D*, i.e. *q* or *not-q*),

$$I(M_i, M_j, e) = \sum_D P(D \mid M_i, e) \log_2 \left(\frac{P(D \mid M_i, e)}{P(D \mid M_j, e)} \right) \tag{5}$$

[3]It appears that using the standard models, this measure will be infinite for the *p* and *not-q* cards, because absolute log-likelihood is infinite when the rule is shown to be false. Therefore, Over and Jessop (1998) assume that rules are defeasible, which means that rules cannot be shown to be false with absolute certainty (see Oaksford & Chater, 1998a, for discussion).

[4]Concerning minimising the number of cards that must be turned over in a sequential version of the task.

Klauer shows that this optimal procedure leads to exactly the same pattern of responses as information gain for the data considered by Oaksford and Chater (1994). Importantly, however, he also shows that these approaches diverge on the effects of prior beliefs, i.e. $P(M_D)$. As for Evans and Over (1996a), Klauer predicts that, if you believe the rule is false, i.e. $P(M_D)$ is low (< .5), then people should select falsifying cases, i.e. they should always select the *not-q* card in preference to the *q* card regardless of $P(p)$ and $P(q)$. This prediction arises because $I(M_i, M_j, e)$ is not symmetrical and so $I(M_I, M_D, e) \neq I(M_D, M_I, e)$. On the optimal procedure used by Klauer, if the $P(M_D) < .5$, i.e. participants do not believe the rule, then the $I(M_I, M_D, e)$ values are used, whereas if $P(M_D) > .05$, i.e. participants do believe the rule, then the $I(M_D, M_I, e)$ values are used. It is this factor that leads the predictions of Klauer's model to diverge from the predictions of the information gain and impact models in the same way as Evans and Over (1996a).

The Effects of Prior Beliefs. We have shown how the various probabilistic accounts diverge on how prior beliefs affect data selection. However, we have concentrated only on how the prior probabilities vary. There have been two other suggestions for how prior beliefs may affect people's data selection performance. First, Oaksford and Chater (1994: see also Oaksford, 1998) suggested that prior beliefs might affect the way people assign probabilities in the task. A constraint on the information gain model is that $P(q) > P(M_D)P(p)$. A problem arises with rules where $P(p)$ is high but $P(q)$ is low (i.e. an "HL" rule), e.g. all black things are ravens. This rule is known to be false: there are many black things that are not ravens. Participants therefore have to decide what to do when asked to test such a rule in the selection task, i.e. they have to reconcile their prior beliefs with the experimental instruction suggesting that it is meaningful to empirically test this rule. Chater and Oaksford (1999a) have argued that it is not always the case that a rule with many exceptions is disbelieved. For example, many people strongly believe that letting children walk home from school increases the likelihood of their being abducted. However, the probability of being abducted given a child walks home from school is very low, i.e. there are many exceptions to this rule. So when participants are confronted by a high $P(p)$ but low $P(q)$ rule to test, they may assume that the experimenter intends them to believe it. Consequently to reconcile a high $P(M_D)$ value with the constraint that $P(q) > P(M_D)P(p)$ requires revising $P(p)$ down. As Oaksford and Chater (1994) suggested, participants may therefore treat high $P(p)$ but low $P(q)$ rules as if they were *low* $P(p)$ and low $P(q)$ rules.

This conjecture has recently been confirmed by Green, Over, and Pyne (1997, see also, Oaksford, 1998). They found that participants' estimates of the probability of finding a *p* on the back of the *not-q* card systematically underestimated $P(p)$ such that $P(p) < P(q)$, even though an HL rule was used,

i.e. $P(p) > P(q)$, and participants' estimates of $P(q)$ were accurate. As Oaksford (1998) points out, this finding is consistent with participants revising down $P(p)$ when $P(p) > P(q)$ as Oaksford and Chater (1994) originally suggested. However, Green and Over (1998) have argued that revising $P(p)$ down is not necessitated by the situation where $P(p) > P(q)$. They observe that, depending on the models that a participant may be considering, it is possible that it would be incoherent to revise $P(p)$ down. We fully agree with Green and Over's (1998) observation. However, we need only claim that revising $P(p)$ down is the normal reaction to this situation and that this is consistent with participants' responses to the HL rule. For example, what would be the appropriate reaction if you were told that *if it's black then it's a raven*? Either you must assume that your interlocutor is uttering totally uninformative falsehoods or that there is an interpretation of their utterance that makes it likely to be true. Perhaps she is referring just to birds in the aviary, where it just so happens that the only black birds are ravens. This amounts to restricting the reference class from *all birds* to *birds in the aviary*. Note also that it means revising $P(p)$ down so that $P(p) \leq P(q)$. (Of course, in this situation $P(q)$ could also change, but with no information to the contrary it should reflect its default rarity value). Consequently we do not dispute Green and Over's (1998) claim that there may be other situations where such a strategy is unreasonable. We would argue, however, that the situation we have described is the usual one and the only one that makes sense of the data. The strategy of revising $P(p)$ down when $P(p) > P(q)$ is exploited in deriving our predictions for the experiments we report here.

A second suggestion for how prior beliefs may affect data selection is that people may compare a hypothesis to more than a single foil, i.e. to more than just M_I (Green & Over, 1997; Over & Jessop, 1998), and that prior beliefs may suggest more appropriate foils, i.e. other than M_I (Green & Over, 1998; Green et al., 1997). For example, suppose you are asked to test the hypothesis that all ravens are pink. Because you know that all ravens are black you might be inclined to test this novel hypothesis against a model where all ravens are not pink, rather than a model that treats whether a bird is a raven or pink as statistically independent, as in Oaksford and Chater (1994).

In our first two experiments we used thematic materials where specific prior beliefs may influence data selection performance. The benefit of such materials is that they avoid constructing complex scenarios that explicitly introduce probabilistic information. These materials therefore avoid cueing participants to the relevance of such information and so we can see if it is used spontaneously. The main potential cost of using such materials is that they may interact with prior beliefs in perhaps unpredictable ways. Anticipating our results, in Experiment 2, it would appear that one rule form triggered a different foil hypothesis, as Over and Jessop (1998) have suggested is a possibility.

Summary

All these probabilistic approaches argue that the *p* card should be selected the most and the *not-p* card selected the least. Consequently these cards cannot discriminate between these approaches. In our analyses we therefore concentrate on the consequent, *q* and *not-q*, cards. We now summarise the key predictions that discriminate between the various theories of the selection task, and which we test in these experiments. These predictions are formulated at general and specific levels, and all involve probabilistic effects:

Prediction 1. Probabilistic vs. *Non-probabilistic Accounts.* Probabilistic approaches predict that probability manipulations will affect card selections, but non-probabilistic accounts do not. This general prediction is confirmed to the extent that our following more specific predictions are confirmed. However, proponents of non-probabilistic accounts could argue that other features of the materials may be responsible for any effects we observe. Consequently in the Discussion we consider each account case by case to see whether any existing distinction made by these theories could account for the effects we observe. Of course proponents of non-probabilistic accounts may be able to formulate ad hoc explanations of any effects we observe but that hardly counts in their favour.

Prediction 2. Kirby (1994) vs. Optimal Data Selection Accounts. Optimal data selection accounts predict (a) more *q* card selections when $P(q)$ is low, and (b) more *not-q* card selections when $P(q)$ is high (see, Oaksford, Chater, Grainger, & Larkin, 1997; Fig. 1). Kirby's (1994) account predicts no such differences. These predictions arise because according to optimal data selection accounts any information that discriminates between M_D and M_I is useful whether it is confirming or falsifying. When $P(q)$ and $P(p)$ are low, it is very surprising to find a *p*, *q* instance and so these are very informative. However, as $P(q)$ or $P(p)$ rise such instances become less surprising and so less informative. Importantly if $P(p)$ is kept constant and $P(q)$ increased then the informativeness of the *q* card falls and that of the *not-q* card rises.

As we discussed in the section Effects of Prior Beliefs, in the selection task people treat high $P(p)$ and low $P(q)$ rules as low $P(p)$ and low $P(q)$ rules. This means that when analysing the selection task data for effects of $P(p)$ we can only treat the high $P(p)$ and high $P(q)$ rule as a genuine high $P(p)$ rule.

Prediction 3. Information Gain and Impact vs. Evans & Over (1996a) and Klauer (1999). Information gain and impact predict that *q* and *not-q* card selections are independent of believability (see Oaksford & Chater, 1994, Fig. 2). In contrast, Evans and Over (1996a) and Klauer (1999) predict that when belief in the rule is low, more *not-q* and fewer *q* cards will be selected. We illustrate how this prediction arises in Fig. 1. In this figure we show the predictions of the various informativeness measures for the case where $P(p)$ and $P(q)$ are both low

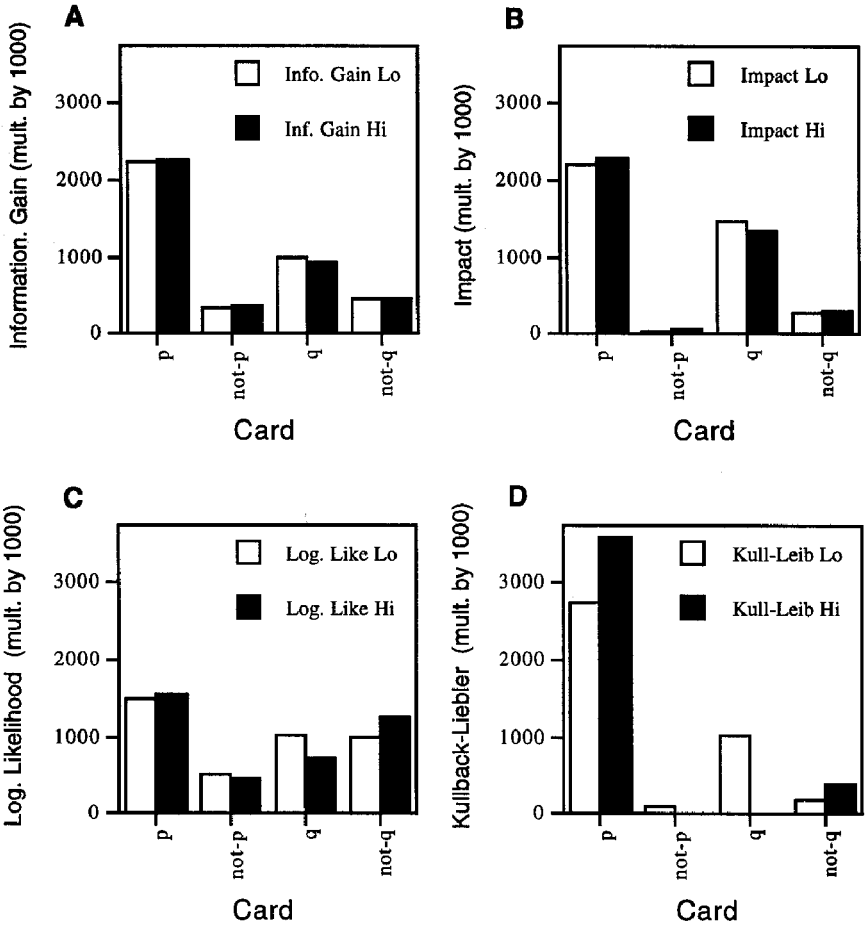

FIG. 1. The predictions of the different probabilistic models of selection task performance, A: Information Gain (Oaksford & Chater, 1994), B: Impact (Nickerson, 1996), C: Log-likelihood ratio (Evans & Over, 1996a); and D: Kullback-Leibler information (Klauer, 1999). The different information measures have been calculated with $P(p) = P(q) = .1$, and at two levels of $P(M_I)$, .4 (Lo) and .6 (Hi). All measures have been scaled by adding .1 to the informativeness of each card and then dividing by the average informativeness, as in Oaksford and Chater (1994) and in Klauer (1999).

but where $P(M_D)$ is varied. Figure 1 shows that when belief in the rule is high, i.e. $P(M_D)$ is high (.6 in Fig. 1) and so $P(M_I)$ is low (.4 in Fig. 1), then all the optimal data selection accounts agree that the *q* card is more informative than the *not-q* card. However, when belief in the rule is low, i.e. $P(M_D)$ is low (.4 in Fig. 1) and so $P(M_I)$ is high (.6 in Fig. 1), then although information gain and impact still predict that the *q* card is more informative than the *not-q* card, Evans and Over (1996a) and Klauer (1999) predict that the *not-q* card is more informative than the *q* card.

EXPERIMENT 1

We investigated these alternative explanations of the selection task by using contents that had been pretested for $P(p)$, $P(q)$, and $P(M_D)$. We fully crossed high and low values of each parameter, creating eight separate rules.

We ensured that each rule was a standard indicative claim about the way the world may be. We describe the rules using ordered couples, $<P(p), P(q)>$, or ordered triples, $<P(p), P(q), P(M_D)>$. Thus "LHH" is the low $P(p)$, high $P(q)$, and high $P(M_D)$ rule and "LH" simply disregards $P(M_D)$, i.e. it refers to both the LHL and the LHH rules. Our strategy will be to check for the predicted effects of the probability manipulations and, should they occur, to dismiss post hoc other possible explanations based on the non-probabilistic approaches described earlier.

There have been other experiments apparently demonstrating effects of varying probabilities in the selection task (Green et al., 1997; Kirby, 1994; Manktelow, Sutherland, & Over, 1995; Oaksford et al., 1997; Pollard & Evans, 1981, 1983). However, Manktelow et al. (1995) investigated the deontic task, not the indicative task that is our current focus. Kirby (1994) showed that increasing the probability of the *p* card increased *not-q* card selections. However, methodological questions about these results have been raised by Over and Evans (1994). Moreover, Kirby did not systematically vary the probability of the *q* card, so his data could not distinguish among the probabilistic approaches. Green et al. (1997) again did not systematically vary $P(p)$ and $P(q)$ in their experiments. Moreover, the interpretation of their results is currently under debate (see Green & Over, 1998; Oaksford, 1998). Importantly, however, Green and Over (1998) recommend that progress in looking for probabilistic effects in the selection task can only be made by obtaining participants' estimates of the relevant probabilities, which is exactly the methodology used in Experiments 1 and 2. Oaksford et al. (1997) demonstrated probabilitistic effects in the reduced array version of the selection task ("RAST"). However, this task version only allows manipulation of $P(q)$ and introduces a sequential sampling element not captured by most of the other theories we introduced earlier (the exception is Klauer, 1999). Moreover, the results in this task have always been aberrant—in an abstract task people appear to make *p*, *not-q* selections. Therefore, judging these accounts on the basis of RAST data may not be considered a fair test.

A number of studies have also varied believability (Fiedler & Hertel, 1994; Love & Kessler, 1995; Pollard & Evans, 1983) and appear to show results consistent with Evans and Over (1996a) and Klauer (1999). However, as Chater and Oaksford (1999a) discuss, these studies did not vary believability directly but only indirectly either by using materials protested by other participants (Pollard & Evans, 1981) or by manipulating the possibility of exceptions rather than believability itself (Fiedler & Hertel, 1994; Love & Kessler, 1995; Pollard &

Evans, 1983).[5] Moreover, none of these studies explicitly obtained measures of participants' degree of belief in the rules they tested in the selection task. Consequently it is difficult to say whether their responses are directly related to their respective degrees of belief. In this experiment we included a probability rating task ("PRT") to obtain estimates of $P(p)$, $P(q)$, and $P(M_D)$ from our participants before or after they performed the selection task. This also served the function of confirming the high–low probability status of the antecedent and consequents of the rules we used.

Finding an HLH rule is difficult because of the problems we discussed in the section *Effects of Prior Knowledge*. However, in protesting materials for this task, we came across an HL rule that was rated by five independent raters as being highly believable. For completeness we therefore included it in our experiment. Of course it is always possible that given a larger sample—128 people participated in this experiment—this rule will not be rated as believable in the PRT.

Method

Participants

A total of 128 undergraduate psychology students from the University of Warwick took part in this experiment. Each participant was paid £4.00 an hour to participate. None of these participants had any prior knowledge of the selection task.

Design

The experiment was a $2 \times 2 \times 2 \times 2$ [$Prob(p) \times Prob(q) \times Prob(M_D) \times$ Order] between-subjects factorial design. Participants were randomly assigned to conditions such that 16 participants performed the experiment with one of the eight task rules. Within each of the eight rule conditions half of the participants received the probability rating task before the selection task and half of the participants received it after the selection task, giving the fourth binary factor.

Materials

The eight rules used in this experiment were as follows (the words in italics correspond to what was seen on the faces of the four cards, the words in parentheses are the corresponding *not-p* or *not-q* instances).

[5] Although the possibility of exceptions and belief in the rule may sometimes go together they are not co-extensive. For example, everyday generalisations like "birds fly" admit of many exceptions (very young birds, injured birds, ostriches, penguins, and so on) but are believed very strongly.

1. If a game is played on a *rink* (*court*) then it is *bowling* (*rugby*). (LLL)
2. If a person is a *politician* (*cleaner*) then they are *privately (state) educated*. (LLH)
3. If a drink is *whisky* (*coffee*) then it is drunk from a *cup* (*glass*). (LHL)
4. If an animal is a *chipmunk* (*cat*) then it has *fur* (*shell*). (LHH)
5. If an item of food is *savoury* (*sweet*) then it is *mousse* (*cheese*). (HLL)
6. If a vegetable is eaten *cooked* (*raw*) then it is a *parsnip* (*cauliflower)*. (HLH)
7. If a flower is *under 1* (*over 1*) *foot tall* then it is domestic (*wild*). (HHL)
8. If an item of furniture is *heavy* (*light*) then it is *big* (*small*). (HHH)

Each rule is a standard indicative claim about how the world is, like the rule in the standard selection task.

Some of the categories used in these rules are binary, for example, *savoury* may be treated as one half of an antonymic pair with *sweet*. However, most of the categories are not binary, for example, *mousse*, *big* (which is one end of a continuum), and *cup*. This division could potentially provide a confounding factor in these experiments. However, Oaksford and Stenning (1992) used binary material in a negations paradigm task and found no deviation from standard selection task performance. The binary material did allow participants to readily identify the contrast set for a negated constituent, e.g. knowing that only two colours, red and blue, were in use allowed participants to interpret "the circle is not blue" unambiguously as "the circle is red". However, in Oaksford and Stenning's (1992) results, binary materials did not lead to any deviations from the standard pattern of results for the rule containing no negations, like all the aforementioned rules. Consequently it seems unlikely that effects observed in the present experiment can be attributed to this aspect of the materials. We will see that Oaksford and Stenning's (1992) results also bear on the interpretation of possible relevance accounts of this experiment, which we discuss later.

The materials consisted of 128 three-page booklets. The first page of each booklet was an instruction page. Depending on the order assigned, the Probability Rating Task (PRT) appeared on one of the following pages and the selection task appeared on the other. In the selection task, for each of the 16 participants in each rule condition, the order in which the materials, i.e. *p*, *not-p*, *q*, and *not-q* instances, appeared on the four cards was randomly selected without replacement from the 24 (4!) possible orders.

Procedure

Participants were tested in three large classroom groups of varying sizes. At the beginning of each class, the booklets were handed out face down and

participants told not to turn them over until instructed. On turning over the booklet the first page revealed the following instructions:

> Your task is to solve the following problems on these pages. You must the problems in order. You may change your mind by crossing an answer out and replacing with another but you may not go back and change your answer once you have turned over the page. Please use whole numbers when responding. Thank you.

The PRT consisted of the following three questions, using the LLH rule as an example:

(Question 1a) Of every 100 people, how many would you expect to be politicians? …

(Question 1b) Of every 100 people, how many would you expect to be privately educated?

> Please estimate on a scale from 0% (must be false) – 100% (must be true) the likelihood that the following statement is true:
>
> (Question 1c) If a person is a politician then they are privately educated.

The instructions for the selection task read as follows, again using the LLH rule as an example:

> Below are four cards. Each card represents a person. One side of each card describes a person's occupation and the other side of each card describes their educational background. Of course you can only see one side of each card.

Below these instructions the four cards were depicted followed by the instruction:

> Your task is to indicate which card or cards you must turn over in order to test that the following rule is true or false:
>
> If a person is a politician then they are privately educated
>
> Please tick those cards you think should be turned over. You may make corrections if you wish, as long as it is unambiguous as to what your final selection is. You may take as long as you like over the problems.

When all participants had finished the booklet they were thanked for their participation. At the end of the experiment participants were fully debriefed concerning the purpose of the experiment.

TABLE 2
Experiment 1: Probability Rating Task

	Low Belief						*High Belief*					
	$P(p)$		$P(q)$		$P(M_D)$		$P(p)$		$P(q)$		$P(M_D)$	
Rule	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*
LL	9.13	8.49	6.38	8.21	19.00	27.74	4.19	9.66	13.06	10.93	34.06	31.45
LH	9.13	15.34	39.31	22.71	14.69	15.08	6.88	11.49	51.44	27.21	85.94	33.63
HL	57.50	17.32	10.63	13.73	12.06	18.94	70.00	24.15	7.00	7.19	12.00	24.92
HH	55.31	24.66	64.38	19.40	34.38	29.15	60.94	23.40	62.19	18.79	68.31	16.56

Mean $P(p)$, $P(q)$, and $P(M_D)$ (in %) for each rule in the probability rating task (PRT) in Experiment 1.

Results and Discussion

Probability Rating Task

Table 2 shows the results of the PRT. Table 2 reveals that with the exception of two $P(M_D)$ values all probability estimates fell within the high–low classification required by the optimal data selection models introduced earlier. $P(p)$ must be greater than approximately .4 to count as a high $P(p)$ rule and $P(q)$ must be greater than approximately .25 to count as a high $P(q)$ rule (see Oaksford & Chater, 1994, Fig. 3). For the LLH and the HLH rules $P(M_D)$ was low, i.e. less than .5. However, as we discussed in the section *Effects of Prior Beliefs*, we did not expect participants to rate HL rules as believable. Because of this result we analyse the effects of believability including and not including the LLH and the HLH rules.

To test whether the task rules affected participants' assessments of $P(p)$, $P(q)$, and $P(M_D)$ as predicted by the pre-test classification we carried out $2 \times 2 \times 2$ [$Prob(p) \times Prob(q) \times Prob(M_D)$] ANOVAs with each of the PRT measures, $P(p)$, $P(q)$, $P(M_D)$, as dependent variables. We then used planned comparisons to check whether the main effect of the corresponding independent variable, $Prob(p)$ $Prob(q)$, $Prob(M_D)$ was significant. The antecedents of the High-$Prob(p)$ rules were rated as more probable than the antecedents of the Low-$Prob(p)$ rules, $F(1, 120) = 286.09$, $MSe = 321.47$, $P < .0001$; the consequents of the High-$Prob(q)$ rules were rated as more probable than the consequents of the Low-$Prob(q)$ rules, $F(1, 120) = 215.46$, $MSe = 301.58$, $P < .0001$; and participants rated the High-$Prob(M_D)$ rules as more believable than the Low-$Prob(M_D)$ rules, $F(1, 120) = 44.28$, $MSe = 652.51$, $P < .0001$. These results confirm the pre-test classification of these rules.[6]

[6]We do not include analyses involving the order in which participants received the tasks, i.e. PRT either before or after the selection task. Although there were some effects of this factor, they were uninterpretable. Importantly we observed the predicted probabilistic effects irrespective of order.

Selection Task

Table 3 shows the percentage of cards selected in the selection task (we discuss the CFI measure below). There are clear effects of the probability manipulations which seem to confirm Prediction 1. We discuss whether the non-probabilistic approach could explain these effects after considering whether the data discriminate between the different probabilistic approaches.

Prediction 1. The probabilistic accounts predict more *not-q* card selections for high *Prob*(p) rules than for low *Prob*(p) rules. To test this prediction we collapsed the data over the high and low *Prob*(M_D) rules. Moreover, as we argued in the section *Effects of Prior Beliefs*, HL rules are treated like LL rules. Therefore we collapsed the HL rule with the LL and LH rules. Significantly more participants selected the *not-q* card when *Prob*(p) was high than when it was low, $\xi^2(1, N = 128) = 8.19$, $P < .005$, consistent with Kirby (1994). This effect was also significant when the HL rules were excluded, $\xi^2(1, N = 96) = 4.67$, $P < .025$, one-tailed.

It could be argued that the PRT indicates that participants did not treat the HL rules as LL rules. Therefore it is illegitimate to collapse the HL rule with the LH and LL rules. However, as we argued in the section *Effects of Prior Beliefs*, the strategy of revising down $P(p)$ is only triggered by the anomaly of being asked to test an HL rule in the selection task phase. Consequently there is no inconsistency in arguing that participants treat HL rules as LL rules in the selection task.

Predictions 1 and 2: Optimal Data Selection vs. Kirby. In contrast to Kirby (1994), optimal data selection approaches predict that *Prob*(q) should affect *not-q* and q card selections (Prediction 2). Significantly more participants selected the *not-q* card when *Prob*(q) was high than when it was low, $\xi^2(1, N = 128) = 6.65$, $P < .005$, one-tailed. Moreover, more participants selected the q card when *Prob*(q) was low than when it was high, albeit not quite significantly, $\xi^2(1, N = 128) = 2.00$, $P = .08$, one-tailed. Nonetheless when *Prob*(p) was low, i.e. for the LL and LH rules, significantly more participants selected the q card when

TABLE 3
Experiment 1: Selection Task

	Low Belief					*High Belief*				
Rule	*p*	*not-p*	*q*	*not-q*	*CFI*	*p*	*not-p*	*q*	*not-q*	*CFI*
LL	81.2	12.5	62.5	18.8	–.44 (.73)	75.0	12.5	50.0	25.0	–.25 (.68)
LH	81.2	12.5	37.5	37.5	.0 (.73)	93.8	0.0	25.0	18.8	–.06 (.57)
HL	87.5	18.8	50.0	6.2	–.44 (.63)	87.5	18.8	62.5	18.8	–.44 (.63)
HH	93.8	43.8	50.0	37.5	–.13 (.72)	87.5	43.8	62.5	56.2	–.06 (.68)

Percentage of cards selected and Mean CFI (sd) for each rule in the selection task in Experiment 1.

Prob(*q*) also was low than when it was high, $\xi^2(1, N = 64) = 4.06, P < .025$. These results are consistent with the optimal data selection accounts, but not with Kirby (1994).

Predictions 1 and 3: Information Gain and Impact vs. Evans and Over (1996a) and Klauer (1999). According to Evans and Over (1996a) and Klauer (1999) there should be more *not-q* card selections when people do not believe the rule. However, there was no significant difference in the frequency of participants selecting the *not-q* card for the low $Prob(M_D)$ rules than for the high $Prob(M_D)$ rules, $\xi^2(1, N = 128) = .35, P = .55$. Moreover, the frequency of participants selecting the *not-q* card for the low $Prob(M_D)$ rules (16/64) was *lower* than for the high $Prob(M_D)$ rules (19/64). It could be argued that the difference was not significant because LLH and HLH are not genuine high $Prob(M_D)$ rules. We therefore excluded these rules and repeated the test, which still did not reach significance, $\xi^2(1, N = 96) = 1.6, P = .10$, one-tailed. Moreover, as before, the proportion of participants selecting the *not-q* card for the low $Prob(M_D)$ rules (16/64) was lower than for the high $Prob(M_D)$ rules (12/32). This effect is in the opposite direction to that predicted by Evans and Over (1996a) or Klauer (1999) and replicates Green and Over's (1997) similar failure to find any effects of believability on card selection. Consequently these results are most consistent with the information gain and impact.[7]

We now consider whether the non-probabilistic approaches could explain the effects we have observed.

PSYCOP and Mental Models. According to both these accounts, response changes in the selection task arise for two reasons. First, the materials alter the balance of people interpreting the rule as a conditional, or as a biconditional. Second, they alter the balance of people who (i) make assumptions about what is on the back of the cards (PSYCOP), or (ii) view their mental model as needing to be "fleshed out". So for example, to explain the significant effect of *Prob*(*p*) on *not-q* card selections, these accounts would have to argue that rules 7 and 8 (see earlier) encourage (i) and (ii) more than rules 1–4. However, current mental logic or mental model theories cannot explain why the HHH rule, *if an item of furniture is heavy then it is big*, should encourage (i) or (ii) any more than the LLH rule, *if a person is a politician then they are privately educated*. A similar argument

[7]Green et al. (1997) presented point-biserial correlations between participants' estimates of *P*(*p*) and whether the *not-q* card was selected. We could also have performed similar analyses using the estimates from the PRT. However, the effects of prior beliefs on this experiment preclude this form of analysis. This is because, as we argued in the section *Effects of Prior Beliefs*, for some rules, e.g. HL, participants re-assess the probabilities when asked to test the rule. Consequently, it is not always the case that the PRT directly reflects the values used in the selection task.

applies to the significant effect of *Prob*(*q*) on *not-q* card selections—why should *if a flower is under 1 foot tall then it is domestic* (HHL) encourage (i) or (ii) any more than *if an item of food is savoury then it is mousse* (HLL). Similarly, to explain why *Prob*(*q*) affects *q* card selections—at least for the low *P*(*p*) rules—these theories would have to explain why *if a person is a politician then they are privately educated* (LLH) leads to more biconditional interpretations than *if an animal is a chipmunk then it has fur* (LHH). Neither Rips' PSYCOP model nor mental models can explain these differences.

Recently, however, Johnson-Laird et al. (in press) have suggested that mental models theory can explain probability judgements in cases where prior knowledge is irrelevant. Moreover, in more realistic settings where prior knowledge is available, Johnson-Laird et al. suggest that individual mental models may be annotated with probabilities, as proposed by Stevenson and Over (1995). Probabilistic calculations over these numbers may explain probabilistic effects in reasoning. Such an account may provide an appropriate algorithmic-level theory of how people reason with conditionals. However, such a theory would have to explain the differences in inferential performance that we observed in this experiment by the probabilistic calculations over the annotations rather than by operations on mental models themselves. Johnson-Laird et al. (in press) do not specify how these calculations are performed but they are clearly not processes similar in kind to the manipulation of mental models. Consequently, there is still no mental model theory that could explain these results.

From a logical point of view, it could be argued that because in Experiment 1 some of the rules are true (4) and some are false (1), it does not make sense to select any data at all. Consequently no predictions can be made about what cards participants should select. Such an objection presupposes a logical approach to data selection explicitly denied by probabilistic approaches. According to these approaches all rules encoding common-sense knowledge are viewed as probabilistic and our belief in them as consequently fallible (Oaksford & Chater, 1998a, in press), i.e. no such rule is believed to be unequivocally true or false. Moreover, probabilistic approaches *do* make explicit predictions about people's data selections when people believe or disbelieve a rule. Consequently, although according to logical approaches it may make no sense to select data to test rules already believed true or false, it does make sense according to probabilistic approaches which do not countenance such extreme commitments.

Pragmatic Reasoning Schema and Darwinian Algorithms. Could our results be explained by participants interpreting some of these rules as deontic regulations or as having a particular cost–benefit structure? Such an interpretation may explain the difference in *not-q* card selections between high *Prob*(*q*) and low *Prob*(*q*) rules. This interpretation would require regarding, for example, the rule *if a flower is under 1 foot tall then it is domestic* (HHL) as a deontic regulation, but not *if an item of food is savoury then it is* mousse (HLL).

However, there is no reason to regard HHL as any "more deontic" than HLL. Moreover, neither can be interpreted deontically because they are both just indicative descriptions of the world and not prescriptions for behaviour. Nonetheless, there can be borderline cases (see Almor & Sloman, 1996). As a test Oaksford and Chater (1996) suggested appending a rule with "It should be the case that". If the resulting sentence makes sense the rule probably has a deontic interpretation. Consider the following:

It should be the case that if you are drinking beer, you are over 21 years of age.
*It should be the case that if a flower is under 1 foot tall then it is domestic.
*It should be the case that if an item of food is savoury then it is mousse.

The first sentence makes sense—the rule is a regulation. However, the second two sentences do not make sense: neither rule *should* be the case, if they are true, they just *are* the case. Consequently these rules are not regulations. Therefore the differences in *not-q* card selections between rules cannot be explained by some being interpreted deontically. Moreover, neither of these rules possesses the cost–benefit structure required to evoke the appropriate Darwinian algorithm, e.g. being a domestic flower cannot be interpreted as a cost paid for the benefit of being under one foot tall.

Relevance and Heuristic Approach. Evans' (1983, 1984, 1989) heuristic approach to relevance cannot explain these results because changes in the balance of card selections, *q* or *not-q*, are only predicted when negations are used in the rules. Therefore the heuristic approach predicts selection of the *p* and *q* card for all rules used in Experiment 1.

Could the differences we observed between rules be explained by Sperber et al.'s (1995) relevance approach? The materials we have used seem to preclude differences in cognitive effects. For the rules where more *not-q* cards were selected, no scenarios were provided indicating that *p*, *not-q* cases were contentious, diagnostic of something that matters, or undesirable (see Sperber et al., 1995, p.61). Consequently whether relevance explains our results depends on whether there were any differences on the effort side. Sperber et al. (1995, p.60) suggest four ways of reducing effort to increase the relevance of the *p*, *not-q* instance. The first three either involve contexts where the *p*, *not-q* instance is made salient or involve a prior learning task, none of which were provided in Experiment 1. Consequently, establishing differential effort relies on showing that the materials introduced a positive feature for the negated categories, the method we discussed in the *Introduction* . However, in Sperber et al. (1995), this was achieved by introducing the relevant features in a brief scenario presented prior to the selection task. Again we did not do this in Experiment 1.

However, taking the HHH rule, the use of "small" on the *not-q* card implicitly provides a positive term for the set of things that can be described as *not-q*, the "*not-q* set". In Sperber et al.'s Experiment 4, for example, in both low effort conditions participants were told that a machine producing cards prints a 4 or a 6 on one side and an A or an E on the other, so that, with reference to the rule *if 6 then E*, *all* the *not-p* cards have a "4" on one side and *all* the *not-q* cards have an "A" on one side. Oaksford and Stenning (1992) refer to the *negated* categories as *contrast classes*. In contrast to Sperber et al.'s Experiment 4, the objects in the *q* card contrast class for the HHH rule do *not* all possess a unique feature, i.e. they are not all small—they may be medium sized, larger than average, quite small, biggish etc. Consequently, the HHH rule does not conform to Sperber et al.'s prescription for how to reduce cognitive effort.

Sperber et al.'s (1995) prescription for reducing cognitive effort suggests that the use of antonymic pairs in the selection task, e.g. sweet/savoury (HLL), should lead to more *not-q* card selections. However, as we pointed out in the *Method* section, Oaksford and Stenning (1992) showed that these materials do not have this effect. Indeed, as they mention, this was clear from the early experiments on the selection task that proposed rules such as *if there is a vowel on one side there is an even number on the other side* (Wason & Johnson-Laird, 1972). "Vowel" is the antonym of "consonant" and "even" is the antonym of "odd". These materials therefore satisfy Sperber et al.'s (1995) prescription for reduced cognitive effort, but they do not lead participants to select the *not-q* card (e.g. Oaksford & Stenning, 1992).[8]

Between-card Differences. Our analyses in Experiment 1 concentrated on between-rule comparisons within a card. However, all the optimal data selection accounts also make certain standard predictions within rules, between cards. So, for example, these accounts predict more *q* than *not-q* card selections for the LL and HL rules (assuming the HL $\Rightarrow$ LL transformation suggested by Oaksford & Chater, 1994), whereas for the LH and HH rules they predict the opposite order. Although in Experiment 1 we did not observe this switch for the LH and HH rules, it is clear from Table 3 that the trends are in the predicted direction—the difference between selections of *q* and *not-q* cards is much smaller for the LH and HH rules than for the LL and HL rules. To test this effect we computed the "consequent falsification index" (CFI, Green et al., 1997; Oaksford et al., 1997; Oaksford & Stenning, 1992) by adding 1 if participants select the *not-q* card and taking 1 away if they select the *q* card. Thus, CFI measures the degree to which

[8]Sperber et al. (1995) do not cite the Oaksford and Stenning (1992) study and so were presumably unaware of this apparent falsification of their account. Although not leading participants to select the *not-q* card, Oaksford and Stenning (1992) do show that such binary materials do remove the matching effect in the negations paradigm selection task (Evans & Lynch, 1973).

participants are aiming to prove the rule false rather than confirming it. We then carried out a 2 × 2 × 2 [$Prob(p) \times Prob(q) \times Prob(M_D)$] ANOVA, with CFI as the dependent variable. If there is a significant trend in the predicted direction then there should be a significant main effect of $Prob(q)$ such that CFI is greater when $Prob(q)$ is high. This was observed, $F(1, 120) = 7.60$, $MSe = .45$, $P < .01$.

Summary

The results of Experiment 1 do not seem consistent with non-probabilistic approaches, and within the probabilistic approaches, they seem most consistent with the information gain and impact measures (Nickerson, 1996; Oaksford & Chater, 1994). However, the support offered for all the probabilistic accounts is equivocal. Although the trends for the consequent cards (*q* and *not-q*) were significant, *not-q* selections did not predominate *q* for the high $P(q)$ rules. This may be because these probabilistic effects are superimposed on some more basic process determining card selection. One possibility is that the *p* and *q* cards were dominant for each rule because of some general default heuristic to pick the possibly confirming instances that applies irrespective of probabilities. This may happen *because* rarity is the norm for indicative rules like those used in Experiment 1, and hence selection of the positive instances has become a hard-wired default option (see Oaksford & Chater, 1994, for discussion). In Experiment 2, we therefore attempted to construct materials that would lead to the predicted pattern of selections within rules as well as between rules.

It could be argued that these rules differ along dimensions other than the probabilities we attempted to manipulate. Consequently any differences may be the result of these factors and not our probabilistic manipulations. Although this is possible, the rules did not differ on any dimension identified as significant by any other theory of selection task performance, as we argued in discussing how these theories might explain these data. Therefore any explanation offered by non-probabilistic accounts would be ad hoc. Nonetheless in Experiment 2 we also sought to remove this potential confound by standardising the contents between rules.

EXPERIMENT 2

In this experiment we used the familiar domain of Members of Parliament ("MPs"). In the UK it is general knowledge that there are a fixed number of MPs, approximately 650. Moreover, rough estimates of the number of MPs from different parties will be common knowledge, e.g. the party forming the Government has a majority in the House of Commons, i.e. more than half of the MPs are in the ruling party.[9] Furthermore, minority parties, e.g. the Ulster

[9]Note that this experiment was conducted before the 1997 UK general election result radically changed the distribution of the parties in the House of Commons.

Unionists or the Scottish Nationalists, are known to have very few seats. We would predict that these familiar contents will overcome any tendency towards a general confirmation bias, if people are attending to the probabilities. Furthermore, these contents allowed us to define rules corresponding to the eight rules used in Experiment 1 but using similar contents about the voting intentions of MPs of different parties. We again included a PRT in order to check that participants' estimates of the various probabilities that an MP is in a certain party and votes in a certain way conformed to the ranges of values demanded by the optimal data selection models.

In this experiment we also attempted to construct HL rules that made more sense to participants than those used in Experiment 1. We did this by varying $P(p)$ and $P(q|not\text{-}p)$ rather than $P(p)$ and $P(q)$. $P(q|not\text{-}p)$ corresponds directly to one of the parameters used in the optimal data selection models. We again used a PRT to gauge the success of this manipulation.

Method

Participants

A total of 80 undergraduate students from the University of Warwick took part in this experiment. Each participant was paid £4.00 per hour to participate. None of these participants had any prior knowledge of the selection task.

Design

The experiment was a 2 × 2 × 2 × 2 [$Prob(p) \times Prob(q) \times Prob(M_D] \times$ Order] mixed design with $Prob(p)$, $Prob(q)$ and Order as between-subjects factors and $Prob(M_D)$ as a within-subject factor. Participants were randomly assigned to conditions such that 20 participants performed the experiment with one of the four types of rule, LL, LH, HL, and HH. Each participant performed two selection tasks, with a low and high believability version of the same rule type. Within each condition, half the participants received the probability rating tasks before the selection tasks and half received them after the selection tasks, leading to the fourth binary Order factor.

Materials

The eight rules used in this experiment were as follows (the words in italics correspond to what was seen on the faces of the four cards, the words in square brackets are the corresponding *not-p* or *not-q* instances).

1. If an MP is a *Scottish Nationalist (MP) [Conservative MP]* then s/he *votes Ulster Unionist [votes Conservative]* in the General Election (LLL)
2. If an MP is a *Scottish Nationalist (MP) [Labour MP}* then s/he *votes Scottish Nationalist [votes Conservative]* in the General Election (LLH)

3. If an MP is a *Scottish Nationalist (MP) [Conservative MP]* then s/he *votes Conservative [votes Labour]* in the General Election (LHL)
4. If an MP is a *Scottish Nationalist (MP) [Conservative MP]* then s/he *votes [abstains]* in the General Election (LHH)
5. If an MP is a *Conservative (MP) [Ulster Unionist MP]* then s/he *abstains* from voting *[votes]* in the General Election (HLL)
6. If an MP is a *Conservative (MP) [Ulster Unionist MP]* then s/he *votes Conservative [votes Labour]* in the General Election (HLH)
7. If an MP is a *Conservative (MP) [Ulster Unionist MP]* then s/he *votes Labour [votes Scottish Nationalist]* in the General Election (HHL)
8. If an MP is a *Conservative (MP) [Ulster Unionist MP]* then s/he *votes [abstains]* in the General Election (HHH)

The materials consisted of 80 five-page booklets similar to those used in Experiment 1. The first page of each booklet was an instruction page. Depending on the order assigned, on the following two pages both Probability Rating Tasks (PRT) appeared and on the remaining two pages the selection tasks appeared, or vice versa.

Procedure

Participants were tested individually. After being seated in the experimental cubicle each participant was given the booklet face down on a table and were told not to turn it over until instructed. On turning over the booklet the first page revealed the same instructions as used in Experiment 1. The PRT was the same form as in Experiment 1.

The instructions for the selection task read as follows, again using the LLH rule as an example:

> A Commons researcher has gathered data about the political allegiance of all Members of Parliament and their voting behaviour. She has recorded the information on a set of cards. Each card represents an MP. One side of each card describes his/her political party and the other side of each card describes their voting intentions in the General Election. Of course you can only see one side of each card.

Below these instructions the four cards were depicted followed by the instruction:

> A reporter suggests that:
>
> If an MP is a Scottish Nationalist the s/he votes Scottish Nationalist in the General Election.
>
> You want to find out whether this is true or false. Before you are four of the researcher's cards. Please tick the card or cards you must turn over in order to test whether the reporter's suggestion is true or false.

You may make corrections if you wish, as long as it is clear what your final selection is. You may take as long as you like over the problems.

When a participant had finished the booklet they were thanked for their participation and were debriefed concerning the purpose of the experiment.

Results and Discussion

Probability Rating Task

Table 4 shows the results of the PRT. Six participants failed to fully complete the PRT and so their data were excluded from the following analyses. Table 4 reveals that, with the exception of $P(q)$ for the HLH rule, all probability estimates fell within the high–low classification required by the optimal data selection models. This indicates that although the HLH rule is now believable (in contrast to Experiment 1), this is because it was treated as an HHH rule. In analysing the PRT we will continue to analyse these data as though they conformed perfectly to our prior high–low classification. We indicate when it is clear that this single discrepancy is responsible for any of the effects we describe.

We used similar analyses as in Experiment 1 to test whether the task rules affected participants assessments of $P(p)$, $P(q)$, and $P(M_D)$. The antecedents of the High-*Prob*(p) rules were rated as more probable than the antecedents of the Low-*Prob*(p) rules, $F(1, 66) = 173.39$, $MSe = 283.95$, $P < .0001$; the consequents of the High-*Prob*(q) rules were rated as more probable than the consequents of the Low-*Prob*(q) rules, $F(1, 66) = 75.64$, $MSe = 653.88$, $P < .0001$; and participants rated the High-*Prob*(M_D) rules as more believable than the Low-*Prob*(M_D) rules, $F(1, 66) = 334.53$, $MSe = 558.11$, $P < .0001$. These results confirm the pre-test classification of these rules.

TABLE 4
Experiment 2: Probability Rating Task

	Low Belief						*High Belief*					
	$P(p)$		$P(q)$		$P(M_D)$		$P(p)$		$P(q)$		$P(M_D)$	
Rule	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*
LL	9.42	10.48	13.05	22.32	9.74	8.89	9.42	10.48	13.84	22.77	88.32	16.17
LH	12.15	15.34	36.00	24.40	22.05	23.43	11.90	15.51	76.10	34.09	76.30	29.59
HL	46.31	9.89	6.63	11.60	9.38	16.49	46.31	9.89	48.44	24.54	85.00	27.77
HH	49.05	12.05	35.63	16.87	5.95	12.11	48.90	13.25	84.63	29.52	81.53	30.69

Mean $P(p)$, $P(q)$, and $P(M_D)$ (in %) for each rule in the probability rating task (PRT) in Experiment 2.

Selection Task

Table 5 shows the percentage of cards selected in the selection task. We first analyse the predictions for the probabilistic approaches to the selection task in the order they were introduced.

Prediction 1: Probabilistic vs. Non-probabilistic Effects. To assess predictions within a card we collapsed the two levels of the within-subject believability factor, $Prob(M_D)$. We performed separate between-subjects 2 × 2 ANOVAs for each card with $Prob(p)$ and $Prob(q)$ as factors and the number of *not-q* card selections as the dependent variable. For the *not-q* card, although no main effects were significant, there was a highly significant interaction between these $Prob(p)$ and $Prob(q)$, $F(1, 70) = 12.04$, $MSe = .54$, $P < .001$, such that *not-q* card selections were high for the LH and HL rules but low for the LL and HH rules. This probabilistic effect is not consistent with the non-probabilistic account of the selection task, thereby confirming Prediction 1.

However, this finding does not appear to be consistent with existing probabilistic models. To examine this possibility we used the $P(p)$, $P(q)$, and $P(M_D)$ values for these rules that we obtained in the PRT to generate information gains according to Oaksford and Chater's (1994) model for each card and found the following correlations with the selection task data: LL rule, $r(6) = .92$, $P < .0025$; LH rule, $r(6) = .92$, $P < .0025$, HL rule, $r(6) = .94$, $P < .001$, HH rule, $r(6) = .11$, *ns.*[10] There was considerable agreement, except for the HH rule. In evaluating Kirby's predictions for the *not-q* card, we therefore excluded the HH rule and contrasted HL with LL and LH collapsed. Consistent with probabilistic accounts there were significantly more *not-q* card selections when $Prob(p)$ was high than when it was low, $F(1, 70) = 6.78$, $MSe = .54$, $P < .025$.

TABLE 5
Experiment 2: Selection Task

	Low Belief					*High Belief*				
Rule	*p*	*not-p*	*q*	*not-q*	*CFI*	*p*	*not-p*	*q*	*not-q*	*CFI*
LL	94.7	5.3	52.6	15.8	–.37 (.68)	100	5.3	63.2	15.8	–.47 (.61)
LH	100	0	45.0	30.0	–.15 (.59)	100	10.0	35.0	50.0	.15 (.49)
HL	93.8	0	25.0	56.2	.31 (.60)	81.2	6.2	25.0	56.2	.31 (.60)
HH	89.5	10.5	47.4	31.6	–.16 (.60)	94.7	5.3	52.6	10.5	–.42 (.61)

Percentage of cards selected and Mean CFI (sd) for each rule in the selection task in Experiment 2.

[10]Note that because we have made the HL rules plausible, it makes sense to test them in the selection task phase without revising $P(p)$.

Participants selection behaviour for the HH rules may be the result of prior beliefs suggesting more appropriate foil hypotheses as suggested by Over and Jessop (1998) and discussed in the section *Effects of Prior Beliefs*. The HHL rule seems to conflict with the reasonable prior belief that MPs of a particular party vote for that party in the general election. This prior belief can be embodied in the probabilistic models as an alternative hypothesis. So rather than comparing the task rule with an independence model, representing no relation between antecedent and consequent, it may instead be compared against a model representing people's more specific prior beliefs. With respect to the HH rules in Experiment 2, people may have strong beliefs that Conservative MPs vote Conservative (HHL), and that the appropriate alternative model for the HHH rule is not independence but where an MP abstains (see Green & Over, 1998, for a related suggestion). In both cases the appropriate foil hypothesis (F) would be the *opposite* model, i.e. one where the dependency is not between p and q (i.e. $P(q|p, M_D)$ is high) but between p and *not-q* (i.e. $P(not\text{-}q|p, M_F)$ is high). As an illustration, we used these models to calculate information gains for the HHL rule using the mean PRT values. Consistent with the results for these rules in Experiment 2, the information gain associated with the q card was almost three times higher than that associated with the *not-q* card. That is, using these models, the q card should still be preferred even though rarity is violated.

Such specific effects of prior beliefs, although perhaps predictable (see section *Effects of Prior Beliefs*), suggest that we should also attempt to demonstrate probabilistic effects using abstract material where specific prior beliefs should not influence the results. Experiments 3 and 4 both used abstract material.

Predictions 1 and 2: Optimal Data Selection vs. Kirby (1994). According to the PRT the HLH rule was treated as an HH rule. We therefore looked initially only at the high $Prob(M_D)$ rules comparing the LLH rule with the LHH and HLH rules collapsed. As predicted, more participants selected the *not-q* card when $Prob(q)$ was high than when it was low, $\xi^2(1, N = 55) = 7.09$, $P < .005$. We then compared the LL and LH rules across both levels of believability by using planned contrasts with the number of *not-q* card selections made as the dependent variable. This comparison was significant in the predicted direction, $F(1, 70) = 4.24$, $MSe = .54$, $P < .05$.

We then performed similar analyses for the q card using the number of q cards selected as the dependent variable. Again there were no significant main effects of $Prob(p)$ or $Prob(q)$. However, there was a significant interaction, $F(1, 70) = 4.76$, $MSe = .71$, $P < .05$, such that q card selections were high for the LL and HH rules but low for the LH and HL rules. We therefore analysed the data as earlier. First we looked only at the high $Prob(M_D)$ rules, comparing the LLH rule with the LHH and HLH collapsed. As predicted more participants selected the q card when $Prob(q)$ was low than when it was high, $\xi^2(1, N = 55) = 5.43$, $P < .01$. We

then compared the LL and LH rules across both levels of believability by using planned contrasts with number of *q* cards selected as the dependent variable. This, however, was not significant, $F(1, 70) = 1.76$, $MSe = .71$, *ns*.

Overall, as in Experiment 1, these data appear most consistent with optimal data selection accounts with respect to Prediction 2.

Predictions 1 and 3: Oaksford and Chater (1994) and Nickerson (1996) vs. Evans and Over (1996a) and Klauer (1999). To assess whether there were more *not-q* card selections when people did not believe the rule we used the McNemar change test (Siegel & Castellan, 1988) which was not significant, $\xi^2(1) = 0$. Indeed the change cell frequencies were identical. This means that exactly the same number of participants selected *not-q* in the high $Prob(M_D)$ condition but not in the low $Prob(M_D)$ condition, as selected this card in the low $Prob(M_D)$ condition but not in the high $Prob(M_D)$ condition. Consequently there was no evidence that people preferred to select the *not-q* card when they did not believe the rule, replicating, within-subject, the results of Experiment 1. Experiments 1 and 2 are the first to provide direct evidence about what the participants actually performing the selection task believe about the rules they are asked to test.

Experiment 2 revealed a variety of probabilistic effects that were predictable only by the probabilistic approaches. That we again found significant effects of $Prob(q)$ on both *q* and *not-q* card selections but failed to find any effects of $Prob(M_D)$ is again most consistent with the information gain and impact accounts.

We now again briefly consider whether the non-probabilistic theories of the selection task could explain the results of Experiment 2.

PSYCOP and Mental Models. In order to explain these data, these accounts must explain why, for example, the LLL rule *If an MP is a Scottish Nationalist then s/he votes Ulster Unionist in the General Election* is interpreted as more biconditional than the LHH rule *If an MP is a Scottish Nationalist then s/he votes in the General Election*, and, moreover, why the LHH rule, but not the LLL rule, is interpreted (i) as requiring assumptions about what is on the back of the cards (PSYCOP), or (ii) as requiring the mental model to be "fleshed out". As for the materials used in Experiment 1 these theories provide no mechanisms to explain why these materials should lead to these different interpretations.

Pragmatic Reasoning Schema and Darwinian Algorithms. Some of the rules used in Experiment 2 may be interpreted deontically. For example, the following seems to make sense:

> It should be the case that if an MP is a Conservative then s/he votes Conservative in the General Election (HLH).

A Conservative MP is obliged to vote Conservative. However, the following seems implausible:

> *It should be the case that if an MP is a Scottish Nationalist then s/he votes Ulster Unionist in the General Election (LLL)

The problem for a deontic interpretation of these results is that the deontic reading makes most sense for the LLH and HLH rule. But these two rules led to different behaviour, e.g. for the LLH rule there were more *q* than *not-q* selections, whereas for the HLH rule the reverse pattern was observed. A similar argument applies in the case of Darwinian algorithms. Although it may make sense to interpret the consequents of the LLH and HLH rules as costs paid (voting for a particular party) for the benefits described in the antecedents (being an MP of that party), the behaviour on these rules was very different. Consequently this pattern of results cannot be explained by participants interpreting some of these rules deontically or in terms of costs and benefits.

Relevance and Heuristic Approaches. Negations were not used in this experiment. Consequently, as for Experiment 1, Evans' (1989) heuristic account does not apply, and hence cannot explain the differences between rules in Experiment 2.

Could Sperber et al.'s (1995) account of relevance explain these results? A brief scenario was introduced in this experiment, but as it was the same for all rules this could not have produced between rule differences on the effect side. On the effort side, only the same prescription discussed in Experiment 1 could apply to these materials, i.e. using antonyms to provide a unique feature for the *not-q* set. However, as we have pointed out, it is questionable whether this manipulation could be primarily responsible for any facilitation effect, because it has been used in experiments where no facilitation is observed (e.g. Oaksford & Stenning, 1992).

Summary

In summary, Experiment 2 produced probabilistic effects that are not consistent with non-probabilistic approaches. Among the probabilistic approaches these data were again most consistent with the information gain and impact approaches. Moreover, the effects for the HH rules could be explained by the probabilistic accounts assuming participants adopted different foils as suggested by Over and Jessop (1998). Note also that in this Experiment, three rules (LHH,

HLL, HLH), led to more *not-q* than *q* card selections. Consequently it would appear that probabilistic manipulations can be powerful enough to override any default confirmation heuristic (Oaksford & Chater, 1994). This result suggests that these probabilistic accounts may not just function as good normative accounts at the computational level, but may also reflect something of the underlying algorithmic processes that may implement these models in the mind.

The precise pattern of results predicted by the probabilistic approaches was not observed because of the effects of specific prior beliefs. In the following two experiments we therefore used abstract materials where such effects should be less in evidence.

EXPERIMENT 3

Sperber et al. (1995) successfully facilitated the logical response in an abstract version of the selection task by manipulating relevance. They constructed experimental materials that they took to vary cognitive effects and effort in the selection task. They argue that their results are consistent with relevance theory, but not other explanations of selection task performance. In contrast, Oaksford and Chater (1995a) argued that probabilistic approaches and relevance accounts are compatible, rather than in competition, and that information gain can be viewed as a quantitative measure of relevance appropriate to the selection task. Oaksford and Chater (1995a) argued for the validity of this interpretation by showing that information gain can explain Sperber et al.'s (1995) experimental results. In particular Oaksford and Chater (1995a) argued that Sperber et al.'s (1995) relevance manipulation using abstract material in their Experiment 4 worked by implicitly manipulating $P(p)$ and $P(q)$. So, for example, in a low effort condition where only two features are used, e.g. the letters printed at random on the cards can only be A or E, implicitly defines a high probability category, i.e. $P(\mathrm{A}) = P(\mathrm{E}) = .5$. If Oaksford and Chater (1995a) are correct, then varying the pattern of high and low effort conditions within rules between the antecedents and consequents should create the LL, LH, HL, and HH rules required to test probabilistic effects. In contrast, Sperber et al. (1995) did not vary effort (and hence according to Oaksford & Chater, 1995a, the probabilities) for antecedent and consequent independently.

We also used a high and a low effects condition in this experiment. If Oaksford and Chater's (1995a) interpretation of Sperber et al.'s (1995) Experiment 4 is correct then only the probability manipulation should influence participants' card selections. In contrast, if Sperber et al. are right about cognitive effects, then we would expect more *not-q* card selections in the high effects condition. Cognitive effects were manipulated as in Sperber et al.'s Experiment 4 by making the *p*, *not-q* instance diagnostic of a fault. However, in their instructions for the high effects condition, Sperber et al. (1995, p.75) explicitly

introduce the *p*, *not-q* instance, "On the back of the card with a 6, the machine has not always printed an E: sometimes it has printed an A instead of an E." This explicit mention of the "6, A" counterexample, means that rather than making the *p*, *not-q* instance relevant, people may be simply "matching" (Evans & Lynch, 1973) the named counterexample to the cards. In this experiment we left out explicit mention of the actual counterexample. If the diagnostic context is what is important, as Sperber et al. (1995) argue, then people should be able to infer the nature of a fault (counterexample) for themselves.

We now contrast the predictions of the two approaches for the present experiment. On the effects side, Sperber et al. (1995) predict that there should be significantly more *not-q* selections in the high rather than the low effects condition. The probabilistic accounts predict no such effect. On the effort side, because effort manipulates probabilities, the probabilistic account makes the same predictions as in Experiments 1 and 2 (Predictions 1–2).

In Experiments 1 and 2, *Prob*(*p*) and *Prob*(*q*) were treated as between-subjects factors. In Experiments 3 and 4 they were treated as within-subject factors.

Method

Participants

A total of 48 undergraduate psychology students from the University of Warwick took part in this experiment. Each participant was paid £4.00 an hour to take part. None of these participants had any prior knowledge of the selection task.

Design

The experiment was a 2 × 2 × 2 [*Prob*(*p*) × *Prob*(*q*) × *Effects*] mixed design with *Prob*(*p*), *Prob*(*q*) as within-subject factors and *Effects* as a between-subjects factor. Participants were randomly assigned to the relevance conditions such that 24 participants performed the experiment in each condition. Each participant performed the four selection tasks with LL, LH, HL, and HH rules presented in random orders.

Materials

In a high effects condition, as in Sperber et al. (1995), it was made clear that the falsifying case was diagnostic of a fault. However, explicit mention of the falsifying case was omitted. As in Kirby (1994), the scenario describes a machine printing cards. In the low *Prob*(*p*) and low *Prob*(*q*), LL, condition the materials were as follows:

A machine manufactures cards.
It is programmed to print at random, on the front of each card, *a letter (A to Z)*.
On the back of each card, it prints a number:
- When there is an A, it prints a 1.
- *When there is a letter B to Z, it prints a number between 1 and 9 (inclusive) at random*.

One day, Mr Jones, the person in charge, realised that the machine has produced some cards it should not have printed.
Mr Jones fixes the machine, examines the newly printed cards and says: don't worry, the machine works fine,

If a card has an A on the front, it has a 1 on the back.

Your task is to indicate which cards need to be turned over in order to establish whether what Mr Jones said is true or false, at least as far as these four cards are concerned. Indicate only the cards that it is absolutely necessary to turn over.

According to Oaksford and Chater (1995a) because *not-p* corresponds to the letters B–Z, and *not-q* corresponds to the numbers 2–9, and these are randomly assigned, $Prob(p) = 1/26$ and $Prob(q) = 1/9$, i.e. they are both low. They also correspond to a high effort condition according to Sperber et al. (1995).

To construct the other probability conditions the clauses in these instructions in italics were replaced with the following. For the LH condition: "a letter (A to Z)" and "When there is a letter B to Z, it prints a 1 or a 2 at random". For the HL condition: "an A or a B" and "When there is a B, it prints a number between 1 and 9 (inclusive) at random". For the HH condition: "an A or a B" and "When there is a B, it prints a 1 or 2 at random".

In the low effect condition, rules and scenarios were constructed that corresponded to Sperber et al.'s (1995) low effect condition. For the low $Prob(p)$ and low $Prob(q)$, LL, condition the materials were as follows:

A machine manufactures cards.
It is programmed to print at random, on the front of each card, *a letter (A to Z)*.
On the back of each card, it prints a number *between 1 and 9 (inclusive)* at random.

The person in charge, Mr Jones, examines the cards and has the strong impression that the machine is not really printing the numbers on the backs at random. He thinks that:

If a card has an A on the front, it has a 1 on the back

but that for the letters B–Z (inclusive) the numbers 1 to 9 are printed on the backs at random. Your task is to indicate which cards need to be turned over in order to establish whether Mr Jones is right or wrong, at least as far as these four cards are concerned. Indicate only the cards that it is absolutely necessary to turn over.

According to Sperber et al. (1995), in these instructions cognitive effects are low because the *p*, *not-q* instance is no longer diagnostic of a fault.

In the low effect condition, to achieve the probability manipulation the clauses in these instructions in italics were replaced with the following. For the LH condition: "a letter (A to Z)" and "1 or 2". For the HL condition: "an A or a B" and "between 1 and 9 (inclusive)". For the HH condition: "an A or a B" and "1 or 2".

In each effects condition the materials consisted of 24 5-page booklets. The first page of each booklet was an instruction page. On the following four pages each of the four selection tasks appeared. Beneath the final line of the instructions on each page four cards were depicted showing *p*, *not-p*, *q*, and *not-q* instances. The order in which these instances appeared on the four cards was randomly selected without replacement from the 24 (4!) possible orders.

Procedure

Participants were tested individually. After being seated in the experimental cubicle each participant was given the booklet face down on a table and was told not to turn it over until instructed. On turning over the booklet the first page revealed the following instructions:

> Please solve the following four problems. There is no time limit so please think carefully before responding.

When all participants had finished the booklet they were thanked for their participation. Participants were debriefed concerning the purpose of the experiment by written note after the final participant had been run in order to avoid communication between participants.

Results and Discussion

Table 6 shows the results of Experiment 3. Analysing the results of Experiments 3 and 4 is simplified because using abstract material means that any probabilistic or relevance effects can only be explained by the probabilistic or relevance models. Thus any such effects count against other approaches. Moreover, we did not manipulate believability in these experiments and so they do not distinguish between the optimal data selection models. We begin as before with the probabilistic approaches.

Probabilistic Approaches

Prediction 1: Probabilistic vs. Non-probabilistic Effects. We tested whether *Prob*(*p*) affects *not-q* card selections by collapsing across the effects conditions. In these analyses we used the proportion of *not-q* cards selected as the dependent variable. The mean proportion of *not-q* card selections when *Prob*(*p*) was high

TABLE 6
Experiment 3

	Low Effects					*High Effects*				
Rule	*p*	*not-p*	*q*	*not-q*	*CFI*	*p*	*not-p*	*q*	*not-q*	*CFI*
LL	70.8	37.5	45.8	45.8	0 (.83)	79.2	4.2	16.7	58.3	.42 (.72)
LH	79.2	41.7	45.8	62.5	.17 (.70)	75.0	8.3	16.7	62.5	.46 (.72)
HL	70.8	37.5	54.2	54.2	0 (.72)	75.0	4.2	16.7	58.3	.42 (.65)
HH	75.0	29.2	54.2	66.7	.13 (.74)	79.2	4.2	8.3	62.5	.54 (.66)

Percentage of cards selected and mean (SD) CFI for each rule in the selection task in Experiment 3.

(mean = .60, sd = .44) did not differ significantly from when *Prob*(*p*) was low (mean = .57, sd = .45), Wilcoxon test, $z = .91$, $P = .18$, one-tailed. This failure to observe a significant increase in *not-q* selections when *Prob*(*p*) was high, may be because for the HL rule with abstract material there is a stronger tendency to revise down *P*(*p*) (Oaksford & Chater, 1994). We therefore compared selections of this card between the LH and HH rules using the McNemar test, which did not approach significance, $\xi^2(1) = .14$, $P = .5$, one-tailed. Consequently there was no evidence in Experiment 3 that *Prob*(*p*) affects the selection of *not-q* cards. This result is not consistent with the probabilistic approaches.

Predictions 1 and 2: Optimal Data Selection vs. Kirby (1994). We performed the same type of analysis for the *not-q* card as described earlier except that we compared the high (LH and HH) and low (LL and HL) *Prob*(*q*) rules. The mean proportion of *not-q* card selections when *Prob*(*q*) was high (mean = .64, sd = .45) was significantly higher than when *Prob*(*q*) was low (mean = .54, sd = .46), Wilcoxon test, $z = 2.07$, $P < .025$, one-tailed. However, the mean proportion of *q* card selections when *Prob*(*q*) was high (mean = .31, sd = .45) was not significantly lower than when *Prob*(*q*) was low (mean = .33, sd = 44), Wilcoxon test, $z = .71$, $P = .24$, one-tailed. Although the result for the *not-q* card is consistent with the optimal data selection approaches, the result for the *q* card is not.

Non-Probabilistic Approaches

The materials and manipulations used in Experiment 3 rule out mental models, mental logics, pragmatic reasoning schemas or Darwinian algorithms as explanations of these data, especially in the light of the very high levels of *not-q* card selections. We therefore concentrate on the relevance approach to explaining these data.

Relevance. According to the relevance approach (Sperber et al., 1995) there should be more *not-q* card selections in the high effects condition than in the low

effects condition. However there was no significant difference in the proportion of *not-q* cards selected between the high (mean = .60, sd = .44) and low (mean = .57, sd = .43) effects conditions, Mann-Whitney test, $z = .13$, $P = .45$, one-tailed. However, a similar analysis did reveal significantly fewer *q* card selections in the high (mean = .15, sd = .31) effects condition that in the low (mean = .57, sd = .43) effects condition, Mann-Whitney test, $z = 2.85$, $P < .005$, one-tailed. Moreover, there were significantly fewer *not-p* card selections in the high (mean = .05, sd = .18) effect conditions that in the low (mean = .37, sd = .38) effects condition, Mann-Whitney test, $z = 3.71$, $P < .0001$, one-tailed.

These results seems to support relevance theory. The high effects condition led to significant reductions in the selection of the non-logical, *not-p* and *q*, cards. This interpretation is further confirmed by the within-rules analysis. We carried out a 2 × 2 × 2 [*Prob*(*p*) × *Prob*(*q*) × *Relevance*] mixed ANOVA, with *Prob*(*p*) and *Prob*(*q*) as within-subject factors and *Relevance* as a between-subjects factor, with CFI as the dependent variable. CFI was significantly higher when *Prob*(*q*) was high than when *Prob*(*q*) was low as predicted by probabilistic approaches, $F(1, 46) = 4.41$, $MSe = .14$, $P < .05$. However, CFI was also higher in the high effects condition than in the low effects condition, $F(1, 46) = 4.46$, $MSe = 1.60$, $P < .05$, which is uniquely predicted by relevance theory.

We also checked for the predicted effects of relevance theory using the frequency of selecting the *p* and *not-q* card combination as the dependent variable. There was a significantly higher proportion of *p*, *not-q* card combinations selected in the high (mean = .45, sd = .43) effects condition than in the low (mean = .23, sd = .41) effects condition, Mann-Whitney test, $z = 1.83$, $P < .05$, one-tailed. There was also a significantly higher proportion of *p*, *not-q* card combinations selected in the high *Prob*(*q*) (mean = .38, sd = .44) condition than in the low *Prob*(*q*) (mean = .30, sd = .45) condition, Wilcoxon test, $z = 1.71$, $P < .05$, one-tailed.

Summary

The pattern of results in Experiment 3 seems most consistent with relevance theory. There was a significant effect of cognitive effects, not predicted by probabilistic accounts. Moreover, it could be argued that the probabilistic effects we observed are also consistent with the relevance account. According to relevance theory low effort conditions facilitate the representation of *p*, *not-q* instances. What matters is being able to identify the contrast class for *not-q* (as suggested by Oaksford & Stenning, 1992). So if only 2 and 7 are possible then a card with *not*-2 on it, must have a 7 on one side. Consequently, according to relevance theory low effort on the consequent is the main requirement for selection of the *p*, *not-q* card combination. The high *Prob*(*q*) categories always correspond to low effort and so relevance can explain the effects of *Prob*(*q*). In our next experiment we therefore kept effort fixed between rules while varying probabilities by other means.

EXPERIMENT 4

Gigerenzer and Hoffrage (1995) showed that many errors and biases in probabilistic reasoning can be removed by using frequency formats, i.e. people will be told that 6 out of 30 swans are white, rather than simply being told that the probability of a swan being white is .2 (see also Gigerenzer, Hell, & Blank, 1988). Such formats seem to encourage people to use probabilistic information appropriately in their reasoning. Recently Oaksford et al. (1997) demonstrated probabilistic effects consistent with optimal data selection using the reduced array version of the selection task. Their version of this task involved showing participants stacks of cards, thereby presenting the probability information concretely in a frequency format. In Experiment 4, therefore, we used a concrete frequency format to manipulate probabilities in the original four card selection task, while using p and q categories designed to create low effort for all rules.

Method

Participants

A total of 20 undergraduate psychology students from the University of Warwick took part in this experiment. Each participant was paid £4.00 per hour to take part. None of these participants had any prior knowledge of the selection task.

Design

The experiment was a 2 × 2 within-subjects design with *Prob*(p), *Prob*(q) as factors. Each participant performed the four selection tasks with LL, LH, HL, and HH rules.

Materials

The materials consisted of two packs of 100 cards each. One pack depicted red circles on one side and the other pack depicted blue triangles on one side. The other side of all the cards in each pack was uniformly patterned. Four stacks of cards were placed in front of participants. One pack contained 10 triangles, one contained 50 circles, one contained 10 blue shapes, and the last contained 50 red shapes. All stacks were placed before participants with the patterned faces uppermost, so that they could not see the coloured shapes on the cards. Each pack had a label behind it. The pack of triangles had a label reading "Triangles", the pack of circles had a label reading "Circles", the pack of blue shapes had a label reading "Blue Shapes", and the pack of red shapes had a label reading "Red Shapes". The rules used in the study described a particular shape as all being of the same colour, e.g. All the triangles are blue. As there were only two shapes and two colours and participants were made aware of this, all the rules in this experiment satisfied Sperber et al.'s prescription for a low effort condition.

Procedure

Participants were tested individually. On entering the experimental room they were seated in front of a table where the experimental materials were laid out. The participant was then given the following instructions to read:

> I have a pack of 120 cards. Each card has a coloured shape on one side. The shapes are either circles or triangles, and the colours are either red or blue. In the pack, 20 of the cards have triangles on them and 100 of the cards have circles on them. 20 of these shapes are blue and 100 of them are red. But you do not know which shapes are which colour.
>
> I have laid out the cards before you so that in one stack there are 10 triangles, in one stack there are 50 circles, in one stack there are 10 blue shapes and in the remaining stack there are 50 red shapes.
>
> Your task is to find out whether *ALL THE TRIANGLES ARE BLUE*. I will shuffle each stack of cards and deal one card from each stack. You must decide which of these four cards (one or more) you must turn over to help find out whether *all the triangles are blue*.

The rule illustrated is the LL rule, i.e. the low *Prob*(*p*) and low *Prob*(*p*) rule. To achieve the probability manipulation, participants were presented with different rules replacing the two occurrences of the rule (in italics) in the instructions. In the LH condition the rule used was "all the triangles are red"; in the HL condition the rule used was "all the circles are blue"; and in the HH condition the rule used was "all the circles are red". The four rule conditions were presented in randomly assigned order. The card dealt from each stack was placed face down in front of the stack from which it was dealt. A new set of instructions containing the new rule was presented in each condition and after each condition the card dealt from each pack was returned ready for the next condition.

Results and Discussion

Table 7 shows the results of Experiment 4. We assess the results in the same way as Experiment 3.

Prediction 1: Kirby (1994)

We tested whether *Prob*(*p*) affects *not-q* card selections using the proportion of *not-q* cards selected for the low *Prob*(*p*) rules (LL and LH) and for the high *Prob*(*p*) rules (HL and HH). The mean proportion of *not-q* card selections when *Prob*(*p*) was high (mean = .48, sd = .41) was significantly lower than when *Prob*(*p*) was low (mean = .65, sd = .37), Wilcoxon test, $z = 1.94$, $P < .05$, one-tailed. This result is significant but in the wrong direction. However, it is only significant because of the HL rule, which as we have noted has peculiar properties. With HL removed there is no significant difference comparing the HH

TABLE 7
Experiment 4

Rule	*P*(*p*)	*P*(*q*)	*p*	*not-p*	*q*	*not-q*	*CFI*
LL	$\frac{1}{6}$	$\frac{1}{6}$	75.0	30.0	65.0	55.0	–.10 (.85)
LH	$\frac{1}{6}$	$\frac{5}{6}$	85.0	10.0	30.0	75.0	.45 (.89)
HL	$\frac{5}{6}$	$\frac{1}{6}$	65.0	20.0	55.0	35.0	–.20 (.89)
HH	$\frac{5}{6}$	$\frac{5}{6}$	75.0	25.0	30.0	60.0	.30 (.87)

Percentage of cards selected and mean (SD) CFI for each rule in the selection task in Experiment 4.

rule with LL and LH collapsed, Wilcoxon test, $z = .54$, $P = .29$, one-tailed. Consequently there was no evidence in Experiment 4 that *Prob*(*p*) affects the selection of *not-q* cards. This result is again not consistent with probabilistic approaches.

Predictions 1 and 2: Probabilistic Approaches

Again we deal with the probabilistic approaches together because they all make the same predictions, that there should be effects of *Prob*(*q*) on *not-q* and *q* card selections. We performed the same type of analysis for the *not-q* card as described earlier except that we compared the high (LH and HH) and low (LL and HL) *Prob*(*q*) rules. The mean proportion of *not-q* card selections when *Prob*(*q*) was high (mean = .68, sd = .41) was significantly higher than when *Prob*(*q*) was low (mean = .45, sd = .39), Wilcoxon test, $z = 2.18$, $P < .025$, one-tailed. Moreover, the mean proportion of *q* card selections when *Prob*(*q*) was high (mean = .30, sd = .41) was significantly lower than when *Prob*(*q*) was low (mean = .60, sd = .35), Wilcoxon test, $z = 3.21$, $P < .001$, one-tailed. These results are consistent with the optimal data selection approaches. They are not consistent with relevance theory because cognitive effort was low for all rules.

Relevance

Every condition in this experiment was a low effort condition, where Sperber et al.'s (1995) relevance account predicts that *not-q* card selections should dominate over *q* card selections. However, in both the LL and HL condition, we found more *q* card than *not-q* card selections as predicted by the optimal data selection models. To test this quantitatively we used CFI as the dependent measure. Contrary to Sperber et al. (1995), the optimal data selection accounts predict a main effect of *Prob*(*q*) such that CFI is lower when *Prob*(*q*) is low than when it is high. We carried out a 2 × 2 ANOVA, with *Prob*(*p*) and *Prob*(*q*) as within-subject factors. Consistent with the optimal data selection accounts but

not with relevance theory, CFI was significantly lower when *Prob*(*q*) was low than when it was high, $F(1, 19) = 13.54$, $MSe = .41$, $P < .0025$. Moreover, for these rules there were no more *p* and *not-q* card combinations (25%) selected than, for example, *p* and *q* card (27.5%) combinations. In sum, only when the effort manipulation is in line with the probability manipulation did we observe the predicted effects of effort. When they are in opposition the probability manipulation dominates. This result is consistent with the view that effort has its effects by manipulating probabilities as Oaksford and Chater (1995a) suggested.

In Experiment 4 the effects of *Prob*(*q*) were consistent with the probabilistic optimal data selection approaches. This experiment replicates for the fourth time the finding that there are more *not-q* card selections when *Prob*(*q*) is high than when it is low, which is uniquely predicted by the optimal data selection approaches to the selection task.

GENERAL DISCUSSION

The purpose of these experiments was to investigate current theories of performance on the indicative selection task where the rules describe how the world is. Experiment 1 used contentful rules whose antecedents and consequents were pre-tested for probability of occurrence. In the selection task the results were most consistent with the information gain and impact theories—card selections were affected by both $P(p)$ *and* $P(q)$ but not by $P(M_D)$.

The materials in Experiment 1 did not lead to more *not-q* than *q* card selections for the LH and HH rules as predicted by the optimal data selection models. In Experiment 2 we used familiar material about the voting behaviour of members of parliament in order to provide a stronger manipulation and to standardise the materials between rules. The results of Experiment 2 were again most consistent with the information gain and impact accounts—card selections were affected by both $P(p)$ *and* $P(q)$ but not by $P(M_D)$. Moreover, with these materials we now observed more *not-q* than *q* card selections, but for the LH and HL rules, and not for the HH rule. This was consistent with participants' PRT results for the HL rules but not for the HH rules.

We argued that these effects may arise as a result of more specific prior beliefs suggesting better foil hypotheses. Consequently in Experiment 3 we moved to abstract materials in order to avoid such effects. Experiment 3 used standard abstract materials and employed the probability manipulations implicit in Sperber et al.'s (1995) effort manipulations and noted by Oaksford and Chater (1995a). We also manipulated effects. We found significant probabilistic effects predicted by the optimal data selection models. However, we also found effects of the effect manipulation. Moreover, the probabilistic effects we observed could be explained by relevance theory.

In Experiment 4 we kept effort constant while manipulating probabilities using a concrete frequency format (Gigerenzer & Hoffrage, 1995) as used by

Oaksford et al. (1997). This experiment confirmed most of the effects predicted by the optimal data selection theories.

Our attempts to manipulate probabilities in these experiments have only been partially successful. No single experiment in this sequence has produced all the effects predicted by probabilistic approaches. The first two experiments attempted to manipulate probabilities using thematic content in order to demonstrate the spontaneous use of probabilistic information. In Experiment 1, although the trends were significantly in the predicted direction, *not-q* card selections did not exceed *q* card selections for the LH or HH rules. In Experiment 2, although *not-q* card selections did exceed *q* card selections for some rules, the HH rules seemed to invoke prior beliefs that led to the standard pattern of card selections where more *q* than *not-q* cards are selected. Nonetheless there are three good reasons to treat these data as supportive of existing probabilistic accounts. First, we did observe significant effects of our probabilistic manipulations. Second, these effects could not be predicted by non-probabilistic accounts. For Experiments 1 and 2 we explored in detail whether non-probabilistic theories could account for these data. Our conclusion was that no existing distinction made by these theories could account for the pattern of effects we observed. These theories may be changed to take account of these data but as we have already pointed out such ad hoc modification is always possible.

Third, the deviations we observed in the data can be explained by theoretical distinctions already in the literature. In Experiment 1 we suggested that people may employ a default confirmation heuristic derived from prior knowledge indicating that such a strategy is the most effective in a world where rarity is the norm. This possibility was first raised by Oaksford and Chater (1994). In Experiment 2 we suggested that where more specific prior beliefs are available people may compare rules to foils other than an independence model. This possibility was raised by Over and Jessop (1998). Consequently, although the theoretical distinctions to which we have appealed have been applied post hoc, they were not developed as an ad hoc response to these data.

It could be argued that although the distinctions to which we appeal are not ad hoc they do amount to allowing probabilistic models a degree of freedom not extended to non-probabilistic models. There are two reasons why our account does not succumb to such an argument. First, such an objection is only possible because the probabilistic models are sufficiently formally precise for the degrees of freedom to be easily specified. For other non-probabilistic accounts it is unclear how many degrees of freedom they already assume. For example, in mental models it appears that the initial model constructed has an undue influence on people's inferential performance (Garnham, 1993; Johnson-Laird & Byrne, 1991). In mental models theory that one model rather than another is chosen as the initial model is treated as a primitive operation (Chater & Oaksford, 1993). However, the processes underlying initial model construction must be very complex and must rely as heavily on prior knowledge as probabilistic

accounts. Consequently, it is unclear whether theories such as mental models do not already assume more degrees of freedom than probabilistic accounts.

Second, in Experiment 3 and 4 we did not use contents that could invoke much prior knowledge and we observed effects that more closely reflected the predictions of the probabilistic models (unmodified by the distinctions invoked in Experiments 1 and 2). However, in Experiments 3 and 4 we also failed to observe some of the effects predicted by probabilistic accounts. In particular we did not observe effects of $P(p)$ and in Experiment 3 most of the effects we observed were consistent with relevance theory (Sperber et al., 1995). As for Experiments 1 and 2 these apparent deviations can be seen as the product of distinctions already drawn by proponents of the probabilistic approach.

Oaksford and Chater (1995a) argued that probabilistic models could be characterised as offering a quantitative account of how prior knowledge affects the relevance of data. That is, probabilistic approaches are not in competition with the relevance account of Sperber et al. (1995). Indeed Oaksford and Stenning (1992) argued that the use of binary materials, which Sperber et al. (1995) use to achieve their effort manipulation, should be interpreted as allowing participants to adopt the most relevant interpretation of a task rule. Consequently we see very few differences between these approaches. Both view the selection task as requiring people to make relevance judgements rather than engage in complex deductive reasoning.[11]

Moreover, it seems likely that the effects manipulation can be explained within decision-theoretic approaches such as Evans and Over (1996a) and Klauer (1999). The effects 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 to explain the results of Green and Over (1997) who manipulated the seriousness of an illness and thereby the costs associated with failing to detect the illness. Of course this suggests that disinterested approaches such as information gain and impact should not be generalised to the effects manipulation. Consequently, these latter approaches are not questioned by the results of Experiment 3.

In Experiments 3 and 4, we did not observe any effects of $P(p)$. A possible reason is that when $P(p)$ is kept constant and $P(q)$ varied, the change in information gain is much higher than when $P(q)$ is kept constant and $P(p)$ varied. For example, in Experiment 4 moving from the LL to the LH rule, i.e. from a low to a high $P(q)$ rule, leads to a 98% increase in the information gain associated with the *not-q* card. However, moving from the LH to the HH rule, i.e. from a low

[11]The main difference appears to concern the role of logic. Sperber et al. (1995) seem to view their account as demonstrating the consistency of selection task behaviour with logic and hence as showing that such results do not question that logic plays a role in human reasoning. In contrast, we regard our probabilistic approach as questioning whether logic plays a significant role in human reasoning.

to a high $P(p)$ rule, leads to only a 5.6% increase in the information gain associated with the *not-q* card.[12] Consequently it would appear that the particular choices of $P(p)$ and $P(q)$ values have led to a situation where the effects of $P(q)$ are much more detectable than those of $P(p)$.

There is recent evidence suggesting that some people may be capable of logical performance on the selection task. According to optimal data selection accounts the logical *p*, *not-q* response only occurs when $P(p)$ or $P(q)$ are high. However, these recent experiments are important because it does not seem plausible to suggest that any manipulation of $P(p)$ or $P(q)$ has occurred to bring about the logical response. Green (1995a,b) has shown that some participants do construe the task logically, and Stanovich and West (1998) have shown that a subgroup (around 10%) of participants with high intelligence are capable of logical performance. These results present few difficulties for probabilistic approaches. We do not deny that the selection task has a logical interpretation: If the rule is interpreted as only applying to the four cards presented in the task, then the task can be construed purely deductively. However, we do deny that this is the most natural interpretation of the task. Interpreting an if … then statement as restricted to a domain of only four objects, is pragmatically odd. For example, suppose there are four men and those who are bald have beards, is it felicitous to describe this situation as "if a man is bald, then he has a beard?" In a directly analogous experiment Legrenzi (1971) found that when presented with the four cards in the selection task so that both sides could be seen, only 1 participant out of 30 described the situation using such a conditional. The most felicitous interpretation of *if a man is bald, then he has a beard* is that it applies to *all* men not just to four. Consequently, the four men should be interpreted as a sample from a larger population and hence a data-selection strategy for verifying universal claims should be adopted. However, some people may interpret the task logically and they may reason about it in a logical way. As Stanovich and West (1998) suggest, if these participants were removed then optimal data selection approaches may reveal even better fits to the data.

More problematic for probabilistic approaches are results that seem to show that in certain circumstances most participants can solve the standard abstract selection task logically. Moshman and Geil (1998) showed that solving the task in groups leads to far higher solution rates. As Stanovich and West (1998) observed, the logical response may emerge because higher IQ participants are able to override their default strategies and think about the task logically. Where a group decision needs to be arrived at, people are likely to take much more time over the task and not simply apply their default strategies. Consequently, this

[12]Only these comparisons made sense due to the problems already discussed at length concerning the HL rule (see section *Effects of Prior Beliefs*). In making these comparisons we assumed that exceptions were possible as in Oaksford and Chater (1998a) and we set the exceptions parameter to .1. $P(M_D)$ was kept at .5 and the $P(p)$ and $P(q)$ values were set from the experimental set up.

situation may permit participants to suspend judgement long enough to see the logical interpretation. It would be interesting to assess the IQs of the individuals in the groups, and to investigate their relative contributions to the collective reasoning process.

Gebauer and Laming (1997) argue that performance in their experiments was logical when participants' interpretations of the task rule are 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. Although this is an interesting result it is inconsistent with the attempts to encourage a logical interpretation in the early literature on the selection task which focused on the same possible misinterpretations (Wason & Johnson-Laird, 1970, 1972). Consequently, these experiments represent an interesting anomaly that deserves further empirical investigation.

In general the reaction of proponents of the probabilistic approach to apparent demonstrations of logicality depends on how thoroughgoing a probabilistic stance is taken. The minimalist approach is to suggest that although a probabilistic optimal data selection account is appropriate for explaining people's normal selection task behaviour, such an account should not be extended elsewhere. This approach may allow that much human reasoning is still deductive in character and should be explained by theories such as mental logic or mental models. Our approach is more thoroughgoing. We suggest that because of the uncertainty of all everyday inference, a probabilistic approach should be adopted to all human reasoning (see Oaksford & Chater, 1998b). For example, we have recently extended the probabilistic approach to syllogistic reasoning (Chater & Oaksford, 1999b) and to conditional inference (Oaksford, Chater, & Larkin, 1998). On our view, apparent displays of logicality in the selection task could arise because restricting the domain of the rule to the four cards may lead participants to ignore prior knowledge. Consequently a probabilistic analysis along the lines proposed by Laming (1996) may be appropriate, which predicts the logical response. Alternatively, it could be conceded that people may have a rudimentary facility for logical thought that, under the right circumstances, can lead to logical performance. Either way logic is denied any central role in everyday human thought.

All the optimal data selection models are defined at the computational level. This contrasts with the non-probabilistic models that typically explain reasoning performance in terms of properties of the algorithms that implement logic in the mind/brain. A legitimate question concerns what process models are suggested by the probabilistic accounts. Oaksford and Chater (1994) suggested that there are a range of possible options from hard-wired heuristics to direct implementations of the mathematical models proposed by these probabilistic theories. They also suggested that a hard-wired heuristic may take the form of a default

confirmation strategy. This would be adaptive because in the normal world where rarity holds it would always select the most informative evidence. However, the current experiments reveal that this could not be the whole story, because such an account would have to predict no behavioural variation as a function of manipulating probabilities in this task. But these experiments show that people are sensitive to probabilistic manipulations. Moreover, our explanation of the aberrant HH rules in Experiment 2 would appear to suggest that people's data selection behaviour may be able to respond flexibly not only to changes in the base rates of the antecedent and consequent but also to differences in relevant prior beliefs—other than their degree of belief in the rule under test [$P(M_D)$]. Oaksford and Chater (1998b) suggest that these probabilistic accounts may be implemented in neural networks following recent proposals that they can be understood as performing various probabilistic computations (Chater, 1995; McClelland, 1998). However, as Marr (1982) argued, the computational level is the starting place for adequate computational models. It is our belief that a great deal more work needs to be done at this level of analysis before considering the cognitive algorithms that implement these accounts. This is because there still seems to be a great deal of scope for explaining many of the observed phenomena purely at the computational level (Oaksford & Chater, 1996).

These results may be relevant to explaining another set of data within the optimal data selection approach. This is the negations paradigm (Evans & Lynch, 1973). In the negations paradigm selection task the antecedent and consequent of a rule can contain negated constituents (*not-p*, *not-q*). There are four possible conditional rules, the original *if p, then q* (AA), together with *if p, then not q* (AN), *if not p, then q* (NA), and *if not p, then not q* (NN). Each participant performs a selection task with each of these four rule types. Using the negations paradigm, Evans and Lynch (1973) reported an effect that they called "matching bias". Participants tend to select the cards that are named in the rules, ignoring the negations. So for example, for the AN rule they will select the p and the q cards, which now correspond to the logical selection, i.e. antecedent, p, and the card representing the denial of the consequent, *not*-(*not-q*) $\Rightarrow q$. Oaksford and Chater (1994) suggested that rather than simply matching, participants are attending to the fact that a negated category is also a high probability category. For example, the probability that you are not drinking whiskey now, is far higher than the probability that you are. This suggests the following equivalences: AA $\Leftrightarrow$ LL; AN $\Leftrightarrow$ LH; NA $\Leftrightarrow$ HL; and NN $\Leftrightarrow$ HH. In the negations paradigm task, participants tend to select the q card throughout and hence make apparently logical selections when the consequent is negated but confirmatory selections when it is not negated. Adopting the equivalence just outlined, our results—especially in Experiment 4—reflect this pattern of responding in task versions where matching could not apply. Consequently, these experiments also provide some support for Oaksford and Chater's (1994) interpretation of the negations paradigm selection task.

Recently Klauer (1999) has shown that the information gain model is sub-optimal when optimality is interpreted as minimising the length of a sequential sample. This demonstration questions whether the information gain model is compatible with rationality. However, the information gain model does provide a close approximation to the optimal decision strategy which is all that can be reasonably expected (Kacelnick, 1998). As Klauer notes, for all the empirical data on the selection task, information gain and the optimal or "Bayes" decision rule make the same predictions. As we have seen they only diverge on the role of prior belief [$P(M_D)$]. Consequently Experiments 1 and 2, which showed that our participants appeared to be insensitive to variation in $P(M_D)$, support the sub-optimal information gain measure (see also Chater & Oaksford, 1999a).

Although the evidence seems to favour the information gain or impact models, Klauer's (1999) analysis may create problems for the claim that information gain provides a rational analysis of the selection task (Oaksford & Chater, 1994, 1996). Following Anderson (1990), Oaksford and Chater (1994) argued that selection task behaviour could be regarded as rational because it reflected an *optimal* adaptation to an environment in which properties are rare. Klauer's analysis implies that if we wish to argue that people are rational because their behaviour is optimal, then showing that behaviour conforms to the information gain model is inadequate because that model is suboptimal. In responding to Klauer's argument, Chater and Oaksford (1999a) argue that information gain and decision-theoretic approaches may apply to different situations. As we discussed in the Introduction, the information gain and impact measures seem most appropriately applied to disinterested inquiry where the costs of making various decisions are not specified. As Chater and Oaksford (1999a) argue, this situation seems to capture the original selection task. However, when appropriate costs are introduced, as we argued occurs in Sperber et al.'s effects manipulation, then a decision-theoretic model may be more appropriate. Consequently with respect its own domain of inquiry, information gain may still be optimal and hence rational (Chater & Oaksford, 1999a).

Manuscript received 26 October 1998
Revised manuscript received 20 January 1999

REFERENCE

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: Lawrence Erlbaum Associates Inc.

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

Braine, M.D.S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, *85*, 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., & Oaksford, M. (1993). Logicism, mental models, and everyday reasoning: Reply to Garnham. *Mind & Language*, *8*, 72–89.

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

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

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.

Cheng, P.W., & Holyoak, K.J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*, 391–416.

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.

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.

Dorling, J. (1983). In philosophical defence of Bayesian rationality. *Behavioral and Brain Sciences*, *6*, 249–250.

Evans, J.St.B.T. (1983). Linguistic determinants of bias in conditional reasoning. *Quarterly Journal of Experimental Psychology*, *35*, 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*. Hove, UK: Lawrence Erlbaum Associates Ltd.

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 in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, *103*, 356–363.

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

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

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

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

Garnham, A. (1993). Is logicist cognitive science possible? *Mind & Language*, *8*, 49–71.

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

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 513–525.

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

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

Green, D.W. (1995b). The abstract selection task: Thesis, antithesis and synthesis. In S. Newstead & J.St.B.T. Evans (Eds.), *Perspective on thinking and reasoning*, (pp. 171–186). Hove, UK: Lawrence Erlbaum Associates Ltd.

Green, D.W., & Larking, R, (1995). The locus of facilitation in the abstract selection task. *Thinking and 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 and Reasoning*, *4*, 231–248.

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

Griggs, R.A., & Cox, J.R. (1983). The effects of problem content and negation on Wason's selection task. *Quarterly Journal of Experimental Psychology*, *35*, 519–533.

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, UK: Cambridge University Press.

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

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

Johnson-Laird, P.N., & Wason, P.C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, *1*, 134–148.

Kacelnick, A. (1998). Normative and descriptive models of decision making: Time discounting and risk sensitivity. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition*. Oxford: Oxford University Press.

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.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological* Review, *94*, 211–228.

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.

Love, R.E., & Kessler, C.L. (l 995). Focusing in Wason's selection task: Content and instruction effects. *Thinking and Reasoning*, *1*, 153–182.

Manktelow, K.I., & Over, D.E. (1993). *Rationality: Psychological and philosophical perspectives*. London: Routledge.

Manktelow, K.I., Sutherland, E.J., & Over, D.E. (1995). Probabilistic factors in deontic reasoning. *Thinking and Reasoning*, *1*, 201–220.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman & Son.

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.

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

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

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

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. (1995a). Information gain explains relevance which explains the selection task. *Cognition*, *57*, 97–108.

Oaksford, M., & Chater, N. (1995b). Theories of reasoning and the computational explanation of everyday inference. *Thinking and 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). 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. (1998b). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning*. Hove, UK: Psychology Press.

Oaksford, M., & Chater, N. (in press). Commonsense reasoning, logic and human rationality. In R. Elio (Ed.), *Commonsense reasoning and rationality*. Oxford: Oxford University Press.

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. (1998). *A probabilistic theory of conditional reasoning*. Unpublished manuscript, School of Psychology, Cardiff University.

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

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

Over, D.E., & Jessop, A.L. (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.

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.

Rips, L.J. (1983). Cognitive processes in propositional reasoning. *Psychological Review*, *90*, 38–71.

Rips, L.J. (1990). Reasoning. *Annual Review of Psychology*, *41*, 321–353.

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

Shannon, C.E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Siegel, S., & Castellan, N.J., Jr. (1998). *Non-parametric 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.

Sperber, D., & Wilson, D. (1986). *Relevance*. Oxford: Basil Blackwell.

Stanovich, K.E., & West, R.F. (1998). Cognitive ability and variation in selection task performance. *Thinking and Reasoning*, *4*, 193–230.

Stevenson, R.J., & Over, D.E. (1995). Deduction from uncertain premises. *Quarterly Journal of Experimental Psychology*, *48A*, 613–643.

Stich, S. (1985). Could man be an irrational animal? *Synthese*, *64*, 115–135.

Stich, S. (1990). *The fragmentation of reason*. Cambridge, MA: MIT Press.

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

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

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.