Data selection and natural sampling: Probabilities do matter

MIKE OAKSFORD and MICHELLE WAKEFIELD
Cardiff University, Cardiff, Wales

Probabilistic accounts of Wason's selection task (Oaksford & Chater, 1994, 1996) are controversial, with some researchers failing to replicate the predicted effects of probability manipulations. This paper reports a single experiment in which participants sampled the data naturally—that is, sequentially. The proportions of possible data types (i.e., cards in the selection task) also reflected the probability manipulation. Other than this procedural difference, the materials were the same as those in Oberauer, Wilhelm, and Rosas-Diaz's (1999) Experiment 3, which failed to show probabilistic effects. Significant probabilistic effects were observed. Moreover, in a comparative model-fitting exercise, a revised version of the information gain model (Hattori, 1999, 2002; Oaksford & Chater, in press-b) was shown to provide better fits to these data than did competing explanations.

Human data selection behavior, as studied in Wason's (1966, 1968) selection task, deviates from the predictions of standard logic. Explanations typically appeal to the way logic is implemented in the mind. Although people are capable of logical reasoning, limited working memory capacity may prevent them from displaying this competence (Johnson-Laird & Byrne, 1991; Rips, 1994). In contrast, it has been suggested that people are not trying to reason logically at all in this task (Oaksford & Chater, 1994, 1996, 1998a, 1998b; Oaksford, Chater, & Grainger, 1999; Oaksford, Chater, Grainger, & Larkin, 1997). Rather, they are reasoning well according to a probabilistic standard. However, despite evidence that probabilities play an important role in data selection behavior (Green & Over, 1997, 2000; Kirby, 1994; Oaksford et al., 1999; Oaksford et al., 1997), this account remains controversial, because some researchers have failed to replicate the predicted effects of probability manipulations (Oberauer, Wilhelm, & Rosas-Diaz, 1999).

The work we report here had three goals. The first goal was to develop a more effective probability manipulation based on *natural sampling* (Gigerenzer & Hoffrage, 1995). The second goal was to test a prediction of a revised version of the *information gain model* (Oaksford & Chater, in press-b) that is not consistent with nonprobabilistic models of the selection task. Nonprobabilistic theories are generally unable to explain why probability manipulations should work. Nonetheless, they may be able to account for the overall pattern of results, because, in these models, participants are assumed to adopt different in-

terpretations of the task. Variations in the number of participants adopting these different interpretations may account for the behavior observed when probabilities are manipulated. Consequently, the third goal was to perform a comparative model-fitting exercise in order to see which model would provide the best overall fit to the data. In a meta-analysis of past results, Oaksford and Chater (in press-b) showed that information gain provided better fits than did other models. It was, therefore, a logical extension to see whether the same would hold for the present data.

We first will introduce the Wason (1966, 1968) selection task. We then will describe the main theoretical approaches in this area. In particular, we will introduce a revised version of the information gain model and show that it makes a unique prediction that is not consistent with nonprobabilistic accounts. The reasons why probabilistic manipulations may have been unsuccessful in the past will then be discussed, and an experimental procedure will be proposed that addresses these problems.

THE SELECTION TASK

In the selection task (Wason, 1966, 1968), an experimenter presents participants with four cards. Each card has a number on one side and a letter on the other. The participants are also presented with a rule of the form *if p then q*—for example, *if there is an A on one side (p), then there is a 2 on the other side (q)*. The four cards show an A (*p* card), a K (*not-p* card), a 2 (*q* card), and a 7 (*not-q* card). The participants have to select those cards that they must turn over to determine whether the rule is true or false. Logically, the participants should select only the *p* and *not-q* cards. However, only 4% of participants make this response, other responses being far more common (*p* and *q* cards, 46%; *p* card only, 33%; *p*,

We thank David Hardman and two anonymous reviewers for their very helpful comments on earlier versions of this paper. Correspondence should be sent to M. Oaksford, School of Psychology, Cardiff University, P.O. Box 901, Cardiff CF10 3YG, Wales (e-mail: oaksford@cardiff.ac.uk).

q , and *not- q* cards, 7%; p and *not- q* cards, 4%; Johnson-Laird & Wason, 1970). All the experiments we discuss used abstract alphanumeric materials, but to introduce the various models and hypotheses, we will use the well-known example of testing the hypothesis that *if a bird is a raven then it is black*.

MODELS OF THE SELECTION TASK

We concentrate on five models of selection task performance, all of which can be parameterized and fit to the data. Two, mental logic and mental models, are non-probabilistic. The rest either are probabilistic or contain a probabilistic component. We will not describe each model in full, since all that is required are the core concepts that allow them to predict the results.

The Revised Information Gain Model

In the information gain model, people select evidence (i.e., turn cards) to distinguish between two hypotheses. In the dependence hypothesis (H_D), whether a bird is black (q) depends on its being a raven (p). In the independence hypothesis (H_I), whether a bird is black or a raven are statistically independent. Contingency tables for each hypothesis are shown in Table 1. In this table, $\varepsilon = P(\text{not-}q|p)$. That is, ε is the probability of an exception—that is, the probability of a nonblack raven. In the normative literature, this probability is called the *conditional uncertainty* (Edgington, 1995). These tables differ from those originally presented by Oaksford and Chater (1994) in two respects. First, the probability of the consequent [$P(q)$] is the same in both hypotheses, which means that looking at nonravens (*not- p* cards) can be informative. Second, the conditional uncertainty [$P(\text{not-}q|p)$] can be nonzero—that is, there may be some nonblack ravens.

Participants are looking for evidence that gives the most information to discriminate between the two hypotheses. As Oaksford and Chater (1996) have shown, calculating expected information gain is equivalent to calculating the *mutual information* (Cover & Thomas, 1991) between the data (what is on the other side of the card) and the hypotheses, given the particular card under consideration. Hattori (1999, 2002) has proposed a logistic selection tendency function that maps this quantity into a probability that a card will be selected. Below, the

equation for the probability of checking ravens to assess their color (i.e., turning the p card) is shown:

$$P(p \text{ card}) = \frac{1}{1 + e^{-2.37 + 9.06 M(p)}}, \quad (1)$$

where

$$M(p) = \frac{I(p)}{\sum_i I(x_i)}, \quad (2)$$

and

$$I(p) = \sum_{ij} P(q_i, H_j | p) \log_2 \left(\frac{P(q_i, H_j | p)}{P(q_i | p) P(H_j | p)} \right). \quad (3)$$

Equation 1 is the selection tendency function. The parameters 2.37 and 9.06 were derived from Hattori (1999). They provide the best fit between the expected information gains and the probability of selecting a card in a meta-analysis of past data. In Equation 2, x_i ranges over the four cards in the task. $M(p)$ is consequently the proportion of the overall information gain available represented by checking a raven for its color (i.e., turning the p card). Equation 3 is the equation for expected mutual information. In Equation 3, q_i ranges over the possible other sides of the p card—for example, whether a raven is black or some other color—and H_j ranges over the two possible hypotheses, H_D and H_I . With some rearrangement, the probabilities in Equation 3 can all be calculated from the contingency tables in Table 1. Assuming that most birds are neither black nor ravens, which we call the *rarity assumption* [i.e., the marginal probabilities $P(p)$ and $P(q)$ are small], it turns out that checking ravens for their color (turning the p card) and black birds for whether they are ravens (turning the q card) is expected to provide the most information about which hypothesis is true. If, on the other hand, these probabilities are high, other cards should be chosen. For example, if most birds are black [i.e., $P(q)$ is high], participants should select the *not- q* card—that is, they should check nonblack birds for whether they are ravens.

The predictions of the revised model are slightly different from those of the original model. In particular, for the probabilities of the antecedent and the consequent used in this experiment, no *crossover* predictions are made such that varying the probability of the antecedent

Table 1
Contingency Tables for the Dependence (H_D) and the Independence (H_I) Hypotheses

	H_D			H_I		
	q	<i>not-q</i>		q	<i>not-q</i>	
p	$a(1 - \varepsilon)$	$a\varepsilon$	a	ab	$a(1 - b)$	a
<i>not-p</i>	$b - a(1 - \varepsilon)$	$(1 - b) - a\varepsilon$	$1 - a$	$b(1 - a)$	$(1 - a)(1 - b)$	$1 - a$
	b	$1 - b$		b	$1 - b$	

Note— a is the probability of p , $P(p)$; b is the probability of q , $P(q)$; and ε is the probability of exceptions, $P(\text{not-}q|p)$.

$[P(p)]$ affects the consequent (q and $not-q$) card selections or varying the probability of the consequent $[P(q)]$ affects the antecedent (p and $not-p$) card selections. For example, keeping $P(q)$ fixed at .9, the predicted probability of selecting the $not-q$ card when $P(p) = .1$ is .09, and it stays at this value, when $P(p) = .9$.¹

Mental Logic and Mental Models

Both the mental models (Johnson-Laird & Byrne, 1991) and the mental logic (e.g., Rips, 1994) accounts of the selection task suggest that only four possible responses can be made on this task. These are (1) p card only, (2) p and q card, (3) p and $not-q$ card, and (4) all four cards. Two processes lead to these different response patterns. First, according to both theories, people interpret the task rule either as a conditional (*if raven then black*) or as a biconditional (*if raven then black and if black then raven*). Second, according to mental models theory, people do not initially fully represent the meaning of these two interpretations. These partial representations lead people to make Selection 1 on a conditional interpretation and Selection 2 on a biconditional interpretation. The same thing happens for the mental logic approach when participants do not consider what could be on the other side of the cards so that they cannot apply backward rules (Rips, 1994). However, if people's mental models are "fleshed out" to fully represent the meaning or people do consider the other sides of the cards, they will make Selection 3 on the conditional interpretation and Selection 4 on the biconditional interpretation.

Validating Search Model

Schroyens and Schaeken's (2003) *validating search model* supplements the mental models theory in the conditional inference paradigm (see Manktelow, 1999, for an overview). After people have constructed an initial representation of a conditional, they perform a validating search of memory for potential counterexamples. So, for example, an initial model may suggest making the *modus ponens* inference (if x is black then x is a raven, a is a raven, therefore a is black). But, rather than just endorsing the conclusion that a is black, people then search long-term memory for a possible counterexample—that is, a nonblack raven. If they find one, they do not make the inference. This model can be applied to the selection task with only minor modifications. In the conditional inference task, the *modus ponens* inference is *not* made if a counterexample is found. However, according to mental models theory, the reverse seems to be true in the selection task. Someone would appear more likely to want to check the color of a particular raven (i.e., turn the p card) if they think they might find that it is not black. Hence, retrieving a counterexample from memory would appear to increase the likelihood that someone will turn a card. So, in contrast to the *modus ponens* inference in the conditional inference task, finding a counterexample in the selection task should incline people to select the p card.

This requires only minor modifications to Schroyens and Schaeken's (2003) equations for the probability that

people will select a card (the expressions in square brackets are their original equations for the conditional inference task):

$$P(p \text{ card}) = CE_{TF} [1 - CE_{TF}]$$

$$P(not-p \text{ card}) = W_{FF} CE_{FT} [W_{FF}(1 - CE_{FT})],$$

and

$$P(q \text{ card}) = CE_{FT} [1 - CE_{FT}]$$

$$P(not-q \text{ card}) = W_{FF} CE_{TF} [W_{FF}(1 - CE_{TF})],$$

where CE_{TF} is the probability of finding a p , $not-q$ counterexample (e.g., a nonblack raven), CE_{FT} is the probability of finding a $not-p$, q counterexample (e.g., a black nonraven), and W_{FF} is the probability of finding a $not-p$, $not-q$ instance² (e.g., a nonblack nonraven). This model, although supplementing mental models theory, introduces a probabilistic component—that is, the probabilities of locating various counterexamples in memory.

Probabilities and Positivity Bias

Although Oberauer et al. (1999) generally failed to find effects of manipulating probabilities in the selection task, they did find some effects in their Experiment 1. They suggested that the participants may select cards in inverse proportion to their frequency. The idea is that rarely occurring cards are just more interesting, irrespective of information gain. This could explain the shift to looking at nonblack birds ($not-q$ cards) when the probability that a bird is black $[P(q)]$ is high: The frequency of nonblack birds ($not-q$ cards) must now be low, and so they will be selected. They also argued that the participants show a positivity bias. That is, they prefer to select the items named in the rule. In all conditions, this is the p and the q cards.

Having summarized the main approaches to the selection task, we will now turn to the novel predictions of the revised information gain model in an experiment in which probabilities were systematically manipulated.

Novel Predictions

Oberauer et al. (1999) manipulated probabilities by informing the participants that the four cards had been selected from a larger pack. They were told, for example, that in a pack of 1,000 cards, 100 have an A on one side, 900 have a K, 900 have a 2, and 100 have a 7. With respect to the task rule *if A then 2*, this is the low-probability antecedent and high-probability consequent rule. In our experiment, this same task rule was used throughout, but the probabilities of the antecedent and the consequent were varied in this way. The different probability conditions were labeled using the ordered pairs $\langle P(p), P(q) \rangle$, with the labels L = low and H = high. So there were four rule types: LL, LH, HL, and HH. After being given this probability information, the participants performed a standard selection task.

The main prediction of the revised information gain model is that a card should be selected more often when the probability of what is on the upturned face is low.

This is a consequence of the rarity assumption and the fact that, for the probabilities used in the present experiment, no crossover predictions are made (see The Revised Information Gain Model section). For example, when the probabilities of ravens (p) and black birds (q) are both low, then according to the model, participants should select ravens (p cards) and black birds (q cards). By raising one of these probabilities, we effectively make the probability of its contrast set—that is, non-ravens or nonblack birds—low. So, by parity of reasoning for the standard LL rule, nonravens or nonblack birds (the *not-p* or *not-q* cards) should now be selected. Consequently, according to the information gain model, ravens (p) should be selected more often when the probability of the antecedent is low—that is, for the LL and LH rules. Conversely, nonravens (*not-p*) should be selected more often when the probability of the antecedent is high—that is, for the HL and HH rules. Black birds (q) should be selected more often when the probability of the consequent is low—that is, for the LL and HL rules. Conversely, nonblack birds (*not-q*) should be selected more often when the probability of the consequent is high—that is, for the LH and HH rules.

None of the nonprobabilistic models directly predicts that probabilistic manipulations should effect card selection performance. Moreover, the revised information gain model makes a prediction that is not consistent with mental models, mental logic, or the validating search model. According to the revised information gain model, when the probabilities of ravens and black birds are both high (for the HH rule), nonravens (*not-p*) and nonblack birds (*not-q*) are more informative than ravens (p) and black birds (q). For example, if the probability of a bird's being a raven is the same as its being black and both are .8 [with $P(H_D) = .5$ and $\varepsilon = .1$], the predicted probability that people check ravens and black birds are both .13, and the predicted probability that people check nonravens and nonblack birds is .84. This means that the participants should select nonravens (*not-p*) more than black birds (q). This is impossible according to the other theories of selection task performance.

According to the mental logic and mental models accounts, a *not-p* card is selected only when people flesh out or consider the other card sides for a *biconditional interpretation*, which will also guide them in selecting the q card. In this case, selections of these two cards must be equal. However, if some participants adopt the biconditional interpretation but do not flesh out, selections of the q card must be higher than selections of the *not-p* card. Consequently, for these models the frequency of q card selections must be greater than or equal to the frequency of *not-p* card selections. A similar relationship holds for the validating search model. The expression for the probability of selecting the *not-p* card is $W_{FF}CE_{FT}$, and the expression for the probability of selecting the q card is CE_{FT} . Because these are probabilities, $W_{FF}CE_{FT}$ must be less than (when $W_{FF} < 1$) or equal to (when $W_{FF} = 1$) CE_{FT} . Consequently, again the frequency of q

card selections must be greater than or equal to the frequency of *not-p* card selections.

We now will discuss some of the problems that may have led to previous failures to replicate probabilistic effects in the selection task. We introduce a new experimental procedure, which we have designed to address these problems and which will allow us to test the predictions of the revised information gain model.

Natural Sampling and Subjective Probabilities

We suggest that failures to replicate probabilistic effects may arise for three main reasons. First, the probability manipulations that have been used (see Green & Over, 1997, 2000; Oaksford et al., 1999; Oaksford et al., 1997; Oberauer et al., 1999) do not match the normal processes by which people update their beliefs about repeatable events in the world. Sampling the real world is a sequential process—that is, each data point is seen one at a time (Gigerenzer & Hoffrage, 1995). During the sampling process, the probabilities, $P(p)$ and $P(q)$, will be learned simultaneously. To reliably detect manipulations away from rarity may require presenting the probability information in the same way that people presumably acquired their default data selection strategy in the first place—that is, by providing each data point one at a time. The experiment we report below had three phases. In the first, the participants were provided with probability information in the same way we introduced above (see the Novel Predictions section). In this phase, they also rated the probabilities of the four possible sides as a check that they had understood the probability manipulation. In the second phase, the participants saw 40 cards, one at a time. The proportions of the different cards reflected the probability manipulation. The participants indicated whether they needed to turn the card over on each trial. As Oaksford et al. (1997) pointed out, using many cards has the potential to introduce confounding strategies. These are discussed in the Results and Discussion section. We now will discuss the motivation for the third phase of the experiment.

A second reason for predictive failures is that experimentally given probabilities have been used to predict the results. However, the cognitive system may have systematically distorted these probabilities, perhaps owing to noisy neural representations. With fixed limits at 0 and 1, error can pull only in one direction—that is, increasing low probabilities but decreasing high probabilities. This pattern of distortion is observed when one compares experimentally given values of $P(p)$ and $P(q)$ and the values of these parameters that provide the best fit to card selection data (Oaksford & Chater, in press-b). To obtain more accurate estimates of the subjective probabilities used in data selection, in the third phase of this experiment, the participants were asked to predict the type of 50 cards card drawn from a pack—that is, a black raven (p, q), a nonblack raven ($p, \text{not-}q$), a black non-raven (*not-p, q*), or a nonblack nonraven (*not-p, not-q*). The proportion of cards identified as being of each type

provided an estimated joint probability distribution for p and q from which $P(p)$, $P(q)$, and ε could be calculated. We predicted that the participants' probability judgments in this phase of the experiment would lead to the pattern of distortion we just introduced, when compared with the initial judgments of these probabilities obtained in Phase 1. Moreover, these probabilities can be used to provide parameter-free fits to the data.

The final reason we considered for predictive failures is that the effects of the rarity assumption have not been taken into account. Some authors have argued that the rarity assumption is made simply to capture the data (Anderson, 1995; Eysenck & Keane, 2000; Laming, 1996). However, it has recently been argued that rarity is the default assumption when people are testing (Anderson & Sheu, 1995; McKenzie & Mikkelsen, 2000) or framing (McKenzie, Ferreira, Mikkelsen, McDermott, & Skrabale, 2001) hypotheses. One consequence is that experimental manipulations that violate rarity are unlikely to totally eliminate the tendency to select the p and q cards. This is because "violations of rarity move the participants' behavior in the appropriate direction but not by as much as the normative theory (unencumbered by strong priors about how the world usually works) would predict" (McKenzie, 2000, p. C7). Thus, probability manipulations are unlikely to lead to behavioral changes as extreme as the model predicts if the experimentally given values are used to derive the predictions of the model (as Oberauer et al., 1999, did). To provide more accurate model predictions therefore requires weighting the probability values toward rarity. We will suggest a way of doing this in the Results and Discussion section.

To provide a rigorous test of the information gain theory, we used a procedure almost identical to that used by Oberauer et al. (1999) in their Experiment 3. In this experiment, the probabilities of the antecedent and consequent of a conditional rule were manipulated to create four different task rules (see the Novel Predictions section). Oberauer et al. found no effects of this probability manipulation.

The HL rule is problematic because most cases are falsifying p , *not- q* cases. It is like being told that *if a bird can fly then it is a raven*. We know this rule is false, because of the many flying birds with which we are familiar that are not ravens (see Oaksford & Chater, 1994, 1998a, for further discussion). However, this does not actually preclude a positive probabilistic dependency between p and q , which may be taken to exist as long as $P(q|p) > P(q)$ —that is, as long as flying (p) and being a raven (q) are not statistically independent. For example, in this condition in the present experiment, the probability of a bird's flying [$P(p)$] is .9, and the probability of its being a raven [$P(q)$] is .1. If participants were to assume that all the ravens can fly, the probability of a bird's being a raven given that it can fly [$P(q|p)$] is .11, and so there is a very weak dependency, so that a bird is more likely to be a raven (q) given that it can fly (p) than you would expect by chance. If this is how the participants view the task for this

rule, there is a clear prediction. Participants' judgments of the conditional uncertainty—that is, $\varepsilon[1 - P(q|p)]$ —should be very high for this condition. This prediction was tested in Phase 3 of the present experiment.

Before describing the experiment, we will briefly reiterate our predictions. First, we would expect the probability assessments in Phase 3 to be systematically distorted, as compared with those obtained in Phase 1, so that low probabilities are overestimated and high probabilities underestimated. Second, if the participants interpret the HL rule as we just outlined, ε should be very high for this rule in Phase 3 of this experiment. Third, for each card, we would expect more to be chosen when the probability of what is on the up-turned face is low. Fourth, for the HH rule, there should be more *not- p* than q card selections. Finally, we would expect the revised information gain model to provide the best overall fits to the data.

METHOD

Participants

Fifty-two undergraduate psychology students from Cardiff University took part in this experiment in return for course credit. None had any prior knowledge of the selection task.

Design

The experiment was a $4 \times 2 \times 2$ mixed design with card (p vs. *not- p* vs. q vs. *not- q*) as a within-subjects factor and $P(p)$ (high [H] vs. low [L]) and $P(q)$ (H vs. L) as between-subjects factors. The two between-subjects factors defined four experimental conditions—that is, LL, LH, HL, and HH.

Materials and Procedure

The experiment was presented on a Macintosh IIsi computer using the PsyScope software (Cohen, MacWhinney, Flatt, & Provost, 1993) to control the presentation of stimuli and to record responses. On arrival in the lab, each participant was randomly assigned to one of the four experimental conditions, so that there were 13 participants in each. The LL condition is used to illustrate the instructions and procedure. On starting the program, the participants were given the following instructions. All the instructions were read at the participants' own pace, and the space bar had to be pressed to initiate the next stage of the experiment.

Thank you for agreeing to participate in this experiment.

The whole procedure should last approximately 20 minutes.

Please take your time to read the instructions carefully before answering any of the questions.

There will be three phases to the experiment and you will be provided with instructions for each phase.

On pressing the space bar, the instructions describing the experimental situation were displayed:

I want you to imagine that you are given a large pile of cards.

The pile consists of 1000 cards and each has a number on one side and a letter on the other side.

Only two letters, A and K, and two numbers, 2 and 7 appear on the cards.

100 of the cards have an A on one side and 900 have a K, and 200 of the cards have a 2 on the other side and 800 have a 7.

In the LH condition, the frequencies for the 2 and 7 cards were replaced with 900 and 100, respectively. For the HL rule, the frequencies were 900 for A, 100 for K, 100 for 2, and 900 for 7. Fi-

nally, for the HH rule, the frequencies were 800 for A, 200 for K, 900 for 2, and 100 for 7. We anchored on .1 as the lowest value and .9 as the highest and added or subtracted .1 to ensure that $P(p) < P(q)$ (apart from the HL rule). This meant that, where possible, the rule could be interpreted as class inclusion, as standard logic requires. When the space bar was pressed, the instructions for the first phase of the experiment were presented. The purpose of this phase was to determine the participants' direct estimates of $P(p)$ and $P(q)$ before they performed the selection task, as in Oaksford et al. (1999, Experiments 1 and 2) and Oberauer et al. (1999, Experiments 2 and 3). The instructions read as follows:

I now want you to imagine that a single card is drawn from the pack of a thousand cards.

What I would like you to do is rate how likely it is on a scale from 0 to 9 that this card has an A, a K, a 2 or a 7 on it.

You will be prompted for each judgment and you should use the number keys at the top of the keyboard to respond.

On pressing the space bar, the participants were presented with four statements, one at a time in random order. Each statement was of the form, "Please rate how likely it is that the card has a K on it," one for each card type, A, K, 2, and 7. When this phase of the experiment was completed, the instructions for the second phase were displayed:

In this second phase of the experiment I want you to test whether the rule:

"if there is an A one side of a card, then there is a 2 on the other side" is true or false of this pack of cards.

To do this I want you to imagine that cards are being drawn at random one at a time from the pile of a 1000 cards.

Each card will be placed in front of you one at a time with only one side showing and your task is to say whether you want to turn the card over to help decide whether the rule is true or false of the pack.

Do not turn cards unnecessarily.

If you want to turn the card, press the Y button on the keyboard.

If you do not want to turn the card, press the N button on the keyboard.

For each card, once you have said whether you want to turn it over or not, you will be asked how confident you are in this decision on a scale from 0 to 9.

Please use the number keys at the top of the keyboard to respond.³

As these instructions reveal, we also collected confidence ratings for each card selection. Reaction times (RTs) were also recorded. On pressing the space bar, the participants were then presented with 40 cards in random order, 20 showing an A or a K and 20 showing a 2 or a 7, in direct proportion to the frequencies specified for each experimental condition. There were 20 letters and 20 numbers, because it was a random binary choice as to which card side was showing on each trial. This procedure provided the participants with direct experience of the probabilities introduced in Phase 1.

Because the cards were not "turned" and, so, the participants remained ignorant of the evidential status of each card, the conditions for applying a sequential decision model (Klauer, 1999) were not met. That is, as in the original task, the participants could not update their beliefs about the rule in response to the data.⁴ This contrasts with the sequential presentation used in the *reduced array* selection task, analyzed by Oaksford et al. (1997).

On completion of this phase, the instructions for the final phase of the experiment were displayed:

In this final phase of the experiment I want you to imagine that I have placed all the cards back into the pack.

I now randomly pick cards out of the pack, BUT you are NOT allowed to see either side of the card.

What I want you to do is assume that the rule is true and then for each card I pick out randomly from the pack, indicate whether you think:

It has an A on one side and a 2 on the other side A2 (PRESS KEY 1), or

It has an 7 on one side and a A on the other side 7A (PRESS KEY 2), or

It has an K on one side and a 7 on the other side K7 (PRESS KEY 3), or

It has an 2 on one side and a K on the other side 2K (PRESS KEY 4) (There are no other possibilities.)

Each time I pick a card out you will be prompted: NEW CARD!

On pressing the space bar, the participants were then prompted 50 times to respond to a "new card's" being drawn from the pack. They were asked to respond as if the rule were true. Therefore, assuming that they probability matched, the frequencies with which they made each response corresponded to the cell values in the joint distribution assuming a dependency between p and q —that is, H_D in Table 1. This allowed $P(p)$, $P(q)$, and the exceptions parameter ε to be calculated. On completion of this phase, a screen was displayed thanking the participants and providing feedback about the purpose of the experiment.

RESULTS AND DISCUSSION

Although we recorded confidence ratings and RTs, we do not report these data, since they do not add much to the analysis of selection frequencies. Confidence ratings and selection frequencies were highly correlated, and RTs were longer for cards with a high proportion of selections.

Probability Estimates

Table 2 shows the probability assessments for each experimental condition when measured directly in Phase 1 and measured indirectly after performing the selection

Table 2
The Direct (Phase 1) and Indirect (Phase 3) Estimates of the Model Parameters in Each Condition of the Experiment

Estimate	Condition	Probability Assessed											
		<i>Prob(p)</i>			<i>Prob(not-p)</i>			<i>Prob(q)</i>			<i>Prob(not-q)</i>		
		<i>M</i>	<i>SD</i>	<i>Med.</i>	<i>M</i>	<i>SD</i>	<i>Med.</i>	<i>M</i>	<i>SD</i>	<i>Med.</i>	<i>M</i>	<i>SD</i>	<i>Med.</i>
Direct	LL	.22	.21	.2	.83	.23	.9	.25	.08	.2	.77	.21	.9
	LH	.49	.32	.3	.74	.20	.8	.80	.22	.9	.23	.20	.1
	HL	.87	.20	.9	.29	.32	.1	.28	.23	.2	.79	.19	.9
	HH	.85	.11	.9	.39	.25	.2	.86	.22	.9	.27	.26	.1
Indirect	LL	.34	.14	.3	.66	.14	.7	.37	.11	.4	.63	.11	.6
	LH	.44	.15	.4	.56	.15	.6	.61	.12	.6	.39	.12	.4
	HL	.68	.17	.6	.32	.17	.4	.36	.17	.4	.64	.17	.6
	HH	.68	.17	.7	.32	.17	.3	.71	.16	.7	.29	.16	.3

Note—*Med.* = median value.

task in Phase 3. Replicating Oberauer et al. (1999), when measured directly, the median values of $P(p)$ and $P(q)$ (to one decimal place) very closely mirrored the given probabilities. However, when measured indirectly in Phase 3, high values were underestimated, and low values were overestimated, as compared with when these probabilities were measured directly in Phase 1.

To quantitatively assess these effects, for each probability judgment— $Prob(p)$, $Prob(not-p)$, $Prob(q)$, and $Prob(not-q)$ —we carried out a 2×2 mixed analysis of variance (ANOVA) with either $P(p)$ or $P(q)$ (L vs. H) as the between-subjects factor and probability estimate (direct vs. indirect) as the within-subjects factor (we use “ $Prob(x)$ ” to distinguish judged probabilities from the experimentally manipulated probabilities). An interaction effect was predicted such that when $P(p)$, $P(not-p)$, $P(q)$, or $P(not-q)$ were high ($>.5$), the indirect estimates should be lower than the direct estimates, whereas when they were low ($\leq .5$), the indirect estimates should be higher than the direct estimates. In all four cases, the interaction was significant [for $Prob(p)$, $F(1,50) = 11.29$, $MS_e = 0.027$, $p < .0025$; for $Prob(not-p)$, $F(1,50) = 5.50$, $MS_e = 0.028$, $p < .025$; for $Prob(q)$, $F(1,50) = 23.20$, $MS_e = 0.019$, $p < .0001$; and for $Prob(not-q)$, $F(1,50) = 16.00$, $MS_e = 0.022$, $p < .0005$]. Simple effects comparisons revealed that for all probability judgments, when the probability was supposed to be high, the indirect estimates were significantly lower than the direct estimates [for $Prob(p)$, $F(1,50) = 16.16$, $MS_e = 0.027$, $p < .0001$; for $Prob(not-p)$, $F(1,50) = 13.86$, $MS_e = 0.029$, $p < .0001$; for $Prob(q)$, $F(1,50) = 19.39$, $MS_e = 0.019$, $p < .0001$; and for $Prob(not-q)$, $F(1,50) = 11.63$, $MS_e = 0.022$, $p < .001$]. However, when the probability was supposed to be low, the indirect estimates were significantly higher than the direct estimates only for $Prob(q)$ [$F(1,50) = 6.03$, $MS_e = 0.019$, $p < .025$] and for $Prob(not-q)$ [$F(1,50) = 5.14$, $MS_e = 0.022$, $p < .05$], but not for the antecedent cards.

The absence of an overestimate effect for the antecedent cards may have been due to some participants’ failing to evaluate the initial probabilities accurately. So even though, for example, they were told that 90% of birds were ravens, they might have rated this probability much lower. We therefore grouped the participants on whether they initially rated the antecedent as high ($>.5$) or low ($\leq .5$) independent of experimental condition. So for example, even though a participant may have been in a high $P(p)$ condition, if he or she rated $P(p)$ as low initially, he or she was assigned to a low $P(p)$ group. Using this grouping as the independent variable revealed all the predicted effects, including the overestimation effect for the antecedent cards.

As we noted in the introduction, the HL rule has always seemed anomalous. Oaksford and Chater (1994) originally argued that people revise down the value of $P(p)$. That is, to make sense of being asked to test the rule *if a bird can fly it is a raven* (see the introduction),

people assume that they are in a context in which flying birds are rare. Other authors have questioned this strategy (e.g., Green & Over, 1998; Green, Over, & Pyne, 1997). As we discussed earlier, an alternative proposal is that people assume a very weak dependency between flying and being a raven. This hypothesis predicts that when confronted by an HL rule, the participants do not revise $P(p)$ down; rather, ε should take on a high value. In the HL condition using the final, indirect estimates, $\varepsilon = .835$. This confirms the prediction of the weak dependency hypothesis and suggests that our original proposal that people revise $P(p)$ down was misguided.

These results show that when the relevant probabilities are assessed indirectly, they follow the pattern observed when experimentally given probabilities are compared with the best-fitting parameter values. Consequently, as Oaksford et al. (1997) argued, part of the reason for why the effects of probability manipulations may not be perfectly captured by the information gain model is a lack of sensitivity to these manipulations. Therefore, it is not legitimate to simply plug in the experimentally given values to the information gain model, as in Oberauer et al. (1999).

Card Selections

Table 3 shows the mean proportions of each card selected in the second, selection task phase of the experiment. To analyze these data, we carried out a $2 \times 2 \times 4$ mixed ANOVA with $P(p)$ (H vs. L) and $P(q)$ (H vs. L) as between-subjects factors and card (p , $not-p$, q , $not-q$) as a within-subjects factor. The only significant main effect was for card [$F(3,144) = 5.37$, $MS_e = 0.14$, $p < .0025$]. This effect was modified by 2 two-way interactions: card $\times P(p)$ [$F(3,144) = 4.25$, $MS_e = 0.14$, $p < .01$] and card $\times P(q)$ [$F(3,144) = 3.68$, $MS_e = 0.14$, $p < .025$]. The three-way interaction was not significant [$F(3,144) = 1.19$]. We further investigated each interaction, using simple effects comparisons.⁵ For the card $\times P(p)$ interaction, there was higher proportion of p cards selected when $P(p)$ was low than when it was high [$F(1,187) = 6.03$, $MS_e = 0.13$, $p < .025$], but there was a higher proportion of $not-p$ cards selected when $P(p)$ was high than when it was low [$F(1,187) = 5.60$, $MS_e = 0.13$, $p < .025$]. There was no effect of $P(p)$ on the consequent cards. For the card $\times P(q)$ interaction, there was a higher propor-

Table 3
The Mean Proportion (and SD) of Each Card Selected in Each Condition of the Experiment

Condition	Card							
	p		$not-p$		q		$not-q$	
	M	SD	M	SD	M	SD	M	SD
LL	.73	.39	.28	.24	.73	.31	.32	.25
LH	.77	.33	.25	.29	.51	.41	.39	.42
HL	.53	.40	.35	.43	.62	.46	.25	.33
HH	.48	.37	.65	.38	.36	.30	.62	.36

tion of q cards selected when $P(q)$ was low than when it was high [$F(1,187) = 5.90$, $MS_e = 0.13$, $p < .025$], but there was a higher proportion of *not- q* cards selected when $P(q)$ was high than when it was low [$F(1,187) = 4.58$, $MS_e = 0.13$, $p < .05$]. There was no effect of $P(q)$ on the antecedent cards.

These results are consistent with the revised information gain model. This is quite strong evidence for this model, since the experimental procedure was very similar to that of Oberauer et al. (1999), who observed no probabilistic effects. Crucially, the prediction that, for the HH rule, the *not- p* card would be selected more often than the q card was significantly confirmed [$t(12) = 2.21$, $p < .025$; see Table 3]. This is again strong evidence for the information gain model, since this finding is not consistent with mental logic, mental models, or the validating search model.

One effect that we did not observe, however, was more *not- q* than q card selections for the LH rule. We argued in the introduction that people's prior knowledge of rarity weights their selection responses (McKenzie et al., 2001). This might be responsible for the failure to see more *not- q* than q card selections for the LH rule. We will explore this possibility when we come to model these data.

As we mentioned in the introduction, having many cards to choose from that occur with different frequencies may introduce confounds with other, more superficial strategies (Oaksford et al., 1997). Random responding, which would predict a 50% response rate for all cards, can be dismissed, because significant differences were found between conditions. However, a strategy whereby a few cards of each type are selected and then no further selections are made could predict the results. This superficial strategy would result in a higher proportion of rarely occurring cards being selected. This means that it could explain the present results, as well as information gain. However, this strategy predicts that the participants should begin by selecting a card but, after a certain number, should stop and not select it again. Across the sequence of cards, this means that the probability of selecting a card should be high initially and then fall to zero.

In contrast, optimal data selection makes a unique prediction for the sequences of individual card selections when $P(p)$ or $P(q)$ is high. When, for example, $P(p)$ is high, the p card should be infrequently selected because of low levels of expected information gain. However, consistent with its normally low default value, people should begin by selecting it, but selections should fall off as they learn that this card is common. This is also consistent with the superficial strategy introduced above. However, this strategy could not predict that when $P(p)$ is high, selections of the low-probability *not- p* card should *rise* over the card sequence. This is because its normal default value is high; therefore, people should begin by not selecting it. Only when they have learned that it is, in fact, rare should they start selecting it. A directly analogous argument applies when $P(q)$ is high.

Before directly testing this hypothesis, we noted that it relies on the assumption that the given probabilities in Phase 1 have had little influence on people's default values. This is just what Oberauer et al. (1999) found using this procedure; that is, they found no effects on card selections of the probability manipulation we used in Phase 1. As is required by our hypothesis, the present experiment replicated Oberauer et al.'s finding. This could be determined because the trial sequence was arranged so that the first four cards included one of each type— p , *not- p* , q , or *not- q* —presented in random order. Consequently we could examine the relationship between people's initial probability assessments and their initial card selection performance. If Phase 1 influenced participants, then for each card, there should be negative point biserial correlations between participants' initial probability ratings and whether or not they selected the card when they first saw it. However, none of these correlations approached significance [p card, $r(50) = -.001$, $p = .49$; *not- p* card, $r(50) = .09$, $p = .28$; q card, $r(50) = -.11$, $p = .23$; *not- q* card, $r(50) = .01$, $p = .48$; all probabilities are one-tailed]. In contrast, similar correlations between the Phase 1 probabilities and the proportion of cards selected in Phase 2 were mostly significant or close to significant [p card, $r(50) = -.21$, $p = .07$; *not- p* card, $r(50) = -.14$, $p = .15$; q card, $r(50) = -.37$, $p = .005$; *not- q* card, $r(50) = -.27$, $p = .03$]. Similar correlations between the Phase 3 probability estimates and the proportion of cards selected were also observed [p card, $r(50) = -.19$, $p = .08$; *not- p* card, $r(50) = -.32$, $p = .01$; q card, $r(50) = -.32$, $p = .01$; *not- q* card, $r(50) = -.42$, $p < .001$]. Thus, although the probability manipulation in Phase 1 had little influence on people's selection task performance, the natural sampling procedure in Phase 2 did affect people's probability estimates and subsequent reasoning in the way predicted.

These analyses establish that people's normal default values were in play at the beginning of Phase 2, despite the probability manipulation in Phase 1. This was an assumption of the hypothesis that we introduced above and that we will now test. When $P(p)$ is high (for the HL and HH rules), the independent variables were card type (p or *not- p*) and card order (first vs. last card experienced). Our hypothesis predicts an interaction such that more p cards should be selected when a card is experienced first than when it is experienced last. In contrast, more *not- p* cards should be selected when a *not- p* card is experienced last than when it is experienced first. A directly analogous prediction holds when $P(q)$ is high (for the LH and HH rules).

We tested this hypothesis, using the sign test. The participants were assigned a score of 1 if they selected a card and of 0 if they did not. When $P(p)$ was high, the proportions of participants selecting cards in each condition were the following: p card, .73 (first), .37 (last); *not- p* card, .42 (first), .63 (last). When $P(q)$ was high, the proportions of participants selecting cards in each condition were the following: q card, .58 (first), .42 (last); *not- q* card, .38 (first), .61 (last). To assess the in-

teractions (a difference between differences), for each participant, a difference score was calculated for each card by subtracting their score for the last card they saw from their score for the first card they saw. When $P(p)$ was high, the difference between these difference scores was significant ($p < .005$). Moreover, both simple effects were significant. More participants selected the p card when they first saw it than when they last saw it ($p < .01$), and more participants selected the *not-p* card when they last saw it than when they first saw it ($p < .05$). When $P(q)$ was high, the difference between these difference scores was also significant ($p = .04$), although neither simple effect reached significance (for both the q and the *not-q* card, $p = .11$). However, the trend for more *not-q* card selections for the card experienced last is not consistent with the suggestion that selections should decrease over the card sequence. These analyses confirm the unique predictions of the revised information gain model and refute the conjecture that the participants' behavior is caused by the superficial strategy we discussed earlier.

Modeling

In this section, we will compare the models of the selection task to see which provides the best overall explanation of these data. Each model's predictions were compared with the mean proportion of cards selected by looking for the parameter values that provided the least sum of squared differences between data and model. This was achieved by using a steepest descent search implemented in Mathematica's MultiStartMin function (Loehle, 2000; Wolfram, 1991). We will report two measures of fit. The coefficient of variation, R^2 , indicates the proportion of variance the model explains and, so, how well it fits the overall trend in the data. We also will report the root mean squared deviation (RMSD), which is related to the average deviation between the predicted value and the observed value of each data point. We fitted the models in two ways. First, as many of the parameters of each model as possible were allowed to vary in order to obtain the best possible fit. Second, when possible, we fixed as many of the parameters of each model to the values obtained in Phase 3 in order to obtain fits to the data with the fewest free parameters as possible. We begin with the information gain model.

Information Gain Model.⁶ For each rule, the information gain model specifies 3 parameters: $P(p)$, $P(q)$, and $\varepsilon [P(\text{not-}q|p)]$. Consequently, the fit to the 16 data points was first calculated with 12 free parameters. However, this procedure led to a region of equal fit. In this region, large variations in ε did not affect the fit and led to only small variations in $P(p)$ and $P(q)$. For example, a change in ε of .9 led to only a .07 change in $P(p)$ and a .09 change in $P(q)$, while leaving the fit unaffected. To present best-fit values, for each rule, ε was therefore fixed to the empirical values obtained in Phase 3 of the experiment. The model was therefore fit with 8 free parameters. The predicted probabilities and the

best-fit parameter values are shown in Appendix A. The fit was very good ($R^2 = .94$)—that is, the model accounted for 94% of the variance in the data. Moreover, the fit to exact location was good (RMSD = .044)—that is, the average deviation between predicted and observed probabilities was .044.

We also fit the model, using the empirical values of the parameters obtained in Phase 3 of the experiment. The effect of the default rarity assumption is to weight the given probabilities toward rarity (McKenzie, 2000; McKenzie et al., 2001). The values of $P(p)$ and $P(q)$ for the LL condition provide the participants' probability estimates when the antecedent and consequent are both rare. Consequently, the effects of rarity can be captured by averaging the values of $P(p)$ and $P(q)$ for the LL condition with the values of these parameters in the other conditions. So, for example, the value of $P(p)$ used to model the HL condition can be calculated as the weighted average of $P(p)$ in the LL and HL conditions. However, we had no theoretical reason for any particular weighting. We therefore treated the weighting as a free parameter but fixed all the other parameters to the empirical values found in Phase 3. We again found good fits to the data. The weighting that produced the best fit was close to .5, with the LL condition being weighted at .48 and the parameters in the remaining conditions being weighted at .52. The information gain model still accounted for 81% of the variance, and the fit to exact location also was good (RMSD = .076). The predicted probabilities of selecting each card are shown in Appendix B.

Mental models and mental logic. Oaksford and Chater (in press-b) have recently presented a simple processing tree implementation (Batchelder & Riefer, 1999) of the mental models (Johnson-Laird & Byrne, 1991) and mental logic (e.g., Rips, 1994) accounts of the selection task. According to these theories, but sticking with the mental models terminology, there are two choice points, each with an associated probability. First, does one adopt a conditional interpretation (P_C) or a biconditional interpretation ($1 - P_C$)? Second, does one flesh out (P_F) or not ($1 - P_F$)? Assuming that these probabilities are independent, a processing tree model yields the following expressions for the probability that each card is selected (where P_E is the probability of error):

$$P(p\text{-card}) = 1 - P_E \quad P(\text{not-}p\text{-card}) = P_F (1 - P_C)$$

and

$$P(q\text{-card}) = 1 - P_C \quad P(\text{not-}q\text{-card}) = P_F.$$

According to these theories, the p card should always be selected, so the only way of allowing that it might not be is to allow for the possibility of error. This means that this model was allowed to fit the data for the p card perfectly, since P_E does not appear in any other expression (Oaksford & Chater, in press-b). Consequently, in fitting this model, exactly the same number of parameters were really free to vary as for the information gain model (i.e., eight). We fitted this model to the data in exactly

the same way as the information gain model (see Appendix A). The fit was not very good ($R^2 = .50$, $\text{RMSD} = .123$). Thus, mental logic and mental models provide poorer fits to the data than does the information gain model. We could not attempt parameter-free fits with this model, since we had no independent data bearing on the parameters involved.

Validating search model. In fitting this model, all 12 parameters (3 for each rule) were allowed to vary. This model provided a much better fit to the data than did the standard mental models account ($R^2 = .89$, $\text{RMSD} = .058$), although still not as good as the information gain model (see Appendix A). Parameter-free fits could be obtained, if it were assumed that the probability of retrieving these instances is the same as the subjective probability that these instances exist. Phase 3 provided data on these latter probabilities, because on this interpretation, $CE_{TF} = P(p, \text{not-}q)$, $CE_{FT} = P(\text{not-}p, q)$, and $W_{FF} = P(\text{not-}p, \text{not-}q)$. However, according to probability theory, $P(p, q) + P(p, \text{not-}q) + P(\text{not-}p, q) + P(\text{not-}p, \text{not-}q) = 1$. But, as is clear from Appendix A, to capture the data, $CE_{TF} + CE_{FT} + W_{FF}$ is greater than 1 for all rules. So under this interpretation, obtaining good fits involves attributing probabilistically inconsistent beliefs to participants (see, also, Oaksford & Chater, 2003, in press-a). Thus, use of the Phase 3 probabilities, which are probabilistically consistent, is bound to lead to very poor parameter-free fits (as indeed it does; $\text{RMSD} = .292$). Moreover, as we have already seen, this model is not consistent with the fact that there were more *not-p* card selections than *q* card selections for the HH rule.

Probabilities and positivity bias. This model predicts that people should select cards in inverse proportion to their probability moderated by positivity bias. Implementing this model is, therefore, straightforward. For each rule, the probability of selecting a card in the selection task is as follows:

$$P(p\text{-card}) = 1 - P(p) \quad P(\text{not-}p\text{-card}) = P(p) - P_b$$

and

$$P(q\text{-card}) = 1 - P(q) \quad P(\text{not-}q\text{-card}) = P(q) - P_b$$

In this model, P_b is the positivity bias parameter, which varies between 0 and 1 and represents how less likely someone is to select a mismatching case. P_b should theoretically take on the same value for all rules, since the difference between matching and mismatching cases is the same across all these rules. The model was therefore fit to the data with nine free parameters. For this model, $R^2 = .94$ and $\text{RMSD} = .044$ —that is, the fit was the same as that with the information gain model, albeit with an extra free parameter.⁷

Phase 3 provided estimates of $P(p)$ and $P(q)$, so parameter-free fits were possible for this model. However, as for the weightings for the information gain model, this experiment provided no independent evidence for any particular value for P_b . Consequently, P_b was free to vary. Its best-fit value was .133. The fit overall was poor ($R^2 = .56$, $\text{RMSD} = .12$). In sum, the infor-

mation gain model provided a better account of these results than did the probabilities and positivity bias model.

CONCLUSIONS

This experiment repeated Oberauer et al.'s (1999) Experiment 3, but with some important modifications. The principal modification involved allowing the participants to experience the relevant probabilities via natural sampling. We also had the participants provide indirect as well as direct estimates of $P(p)$ and $P(q)$. A comparison of these estimates showed that the participants overestimated the low probabilities and underestimated the high probabilities. Contrary to Oberauer et al., we found strong effects of the probability manipulations, consistent with the information gain model, but not with any other theory of selection task performance. The finding of more *not-p* card selections than *q* card selections for the HH rule was not consistent with the mental models, the mental logic, or the validating search model. The mental models and mental logic approaches also provided poor overall fits to the data. The validating search model provided a much better overall fit, but only at the cost of attributing probabilistically inconsistent beliefs to people. The probabilities and positivity bias model captured the HH rule effect and provided similar overall fits. However, the fit with as few free parameters as possible was poor. In sum, these data show that probabilities do matter in the selection task and that the revised information gain model provides the best overall explanation of the effects observed when probabilities are varied in this task.

REFERENCES

- ANDERSON, J. R. (1995). *Cognitive psychology and its implications*. New York: Freeman.
- ANDERSON, J. R., & SHEU, C.-F. (1995). Causal inferences as perceptual judgments. *Memory & Cognition*, **23**, 510-524.
- BATCHELDER, W. H., & RIEFER, D. M. (1999). Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin & Review*, **6**, 57-86.
- COHEN, J., MACWHINNEY, B., FLATT, M., & PROVOST, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers*, **25**, 257-271.
- COVER, T. M., & THOMAS, J. A. (1991). *Elements of information theory*. New York: Wiley.
- EDGINGTON, D. (1995). On conditionals. *Mind*, **104**, 245-329.
- EYSENCK, M. W., & KEANE, M. T. (2000). *Cognitive psychology*. Hove, U.K.: Psychology Press.
- GIGERENZER, G., & HOFFRAGE, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, **102**, 684-704.
- GREEN, D. W., & OVER, D. E. (1997). Causal inference, contingency tables and the selection task. *Current Psychology of Cognition*, **16**, 459-487.
- GREEN, D. W., & OVER, D. E. (1998). Reaching a decision: A reply to Oaksford. *Thinking & Reasoning*, **4**, 231-248.
- GREEN, D. W., & OVER, D. E. (2000). Decision theoretical effects in testing a causal conditional. *Current Psychology of Cognition*, **19**, 51-68.
- GREEN, D. W., OVER, D. E., & PYNE, R. A. (1997). Probability and choice in the selection task. *Thinking & Reasoning*, **3**, 209-236.

- HATTORI, M. (1999). The effects of probabilistic information in Wason's selection task: An analysis of strategy based on the ODS model. In *Proceedings of the 16th Annual Meeting of the Japanese Cognitive Science Society*, **16**, 623-626.
- HATTORI, M. (2002). A quantitative model of optimal data selection in Wason's selection task. *Quarterly Journal of Experimental Psychology*, **55A**, 1241-1272.
- JOHNSON-LAIRD, P. N., & BYRNE, R. M. J. (1991). *Deduction*. Hove, U.K.: Erlbaum.
- JOHNSON-LAIRD, P. N., & WASON, P. C. (1970). Insight into a logical relation. *Quarterly Journal of Experimental Psychology*, **22**, 49-61.
- 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.
- LAMING, D. (1996). On the analysis of irrational data selection: A critique of Oaksford and Chater (1994). *Psychological Review*, **103**, 364-373.
- LOEHLE, C. (2000). *Global Optimization 4.0* [Computer program]. Naperville, IL: Loehle Enterprises.
- MANKTELOW, K. I. (1999). *Reasoning and thinking*. Hove, U.K.: Psychology Press.
- MCKENZIE, C. R. M. (2000). *Examining the rarity assumption and its implications*. Unpublished manuscript, University of California, San Diego, Department of Psychology.
- MCKENZIE, C. R. M., FERREIRA, V. S., MIKKELSEN, L. A., McDERMOTT, K. J., & SKRABLE, R. P. (2001). Do conditional statements target rare events? *Organizational Behavior & Human Decision Processes*, **85**, 291-309.
- MCKENZIE, C. R. M., & MIKKELSEN, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin & Review*, **7**, 360-366.
- 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. (1996). Rational explanation of the selection task. *Psychological Review*, **103**, 381-391.
- OAKSFORD, M., & CHATER, N. (1998a). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning*. Hove, U.K.: Psychology Press.
- OAKSFORD, M., & CHATER, N. (1998b). A revised rational analysis of the selection task: Exceptions and sequential sampling. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 372-398). Oxford: Oxford University Press.
- OAKSFORD, M., & CHATER, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, **5**, 349-357.
- OAKSFORD, M., & CHATER, N. (2003). Computational levels and conditional reasoning: A reply to Schroyens and Schaeken. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 150-156.
- OAKSFORD, M., & CHATER, N. (in press-a). Conditional probability and the cognitive science of conditional reasoning. *Mind & Language*.
- OAKSFORD, M., & CHATER, N. (in press-b). Optimal data selection: Revision, review and re-evaluation. *Psychonomic Bulletin & Review*.
- OAKSFORD, M., CHATER, N., & GRAINGER, B. (1999). Probabilistic effects in data selection. *Thinking & Reasoning*, **5**, 193-243.
- OAKSFORD, M., CHATER, N., GRAINGER, B., & LARKIN, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 441-458.
- OBERAUER, K., WILHELM, O., & ROSAS-DIAZ, R. (1999). Bayesian rationality for the Wason selection task? A test of optimal data selection theory. *Thinking & Reasoning*, **5**, 115-144.
- RIPS, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.
- SCHROYENS, W., & SCHAEKEN, W. (2003). A critique of Oaksford, Chater, and Larkin's (2000) conditional probability model of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 140-149.
- WASON, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology* (pp. 135-151). Harmondsworth, U.K.: Penguin.
- WASON, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, **20**, 273-281.
- WINER, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.
- WOLFRAM, S. (1991). *Mathematica* [Computer program]. Reading, MA: Addison-Wesley.

NOTES

1. However, keeping $P(q)$ fixed at .6, the predicted probability of selecting the *not-q* card when $P(p) = .1$ is .29, but when $P(p) = .5$, this probability goes up to .74. So the model still makes crossover predictions.

2. Schroyens and Schaeken (2003) argued that people will endorse inferences such as denying the antecedent (if p then q , *not-p*, therefore *not-q*) only if they can retrieve some *not-p*, *not-p* instances (i.e., W_{FF}); that is, the higher this probability, the more likely you are to draw this inference. A similar argument would seem to apply to the selection task. These are confirming instances, so the more someone can retrieve, the more likely he or she is to select the card.

3. It could be argued that asking the participants to turn the cards they "want" to turn versus turning the cards they *must* turn somehow biases them to interpret the task inductively rather than deductively. However, the participants were also told, "Do not turn cards unnecessarily." This instruction means to turn only those cards it is necessary to turn, or in paraphrase, to turn only the cards they must turn. Consequently, the fact that "want" was used cannot be argued to bias people away from a possible deductive interpretation of the task. Moreover, the participants were testing whether a sample provides evidence for whether the rule is true or false of the pack. Since they did not get to see the rest of the pack, deductive certainty was impossible. Oberauer et al. (1999) did not treat this as ruling out deductive approaches, and rightly so. If people adopted a Popperian view, the failure to observe falsifying instances would "corroborate" the hypothesis about the pack. Thus, if the participants are natural falsificationists about testing hypotheses, they should "want" to look at potentially falsifying cases.

4. By using the original hypotheses in Oaksford and Chater (1994), this procedure could have influenced the participants' belief in the rule, because $P(q)$ varied between models, so that it was more likely in H_D than in H_I . However, $P(q)$ is the same in both hypotheses in the revised information gain model (Oaksford & Chater, 2001, in press-b), so simply observing the frequency of q now provides no information about which hypothesis is true.

5. Throughout this paper, when simple effects were computed, the pooled error term was used, and Satterwaite's correction was applied to calculate the degrees of freedom (see Winer, 1971, pp. 529-532).

6. In this experiment, the participants were made aware at the outset that there were four types of information that they must choose between. That is, they knew that cards could have only As, Ks, 2s, and 7s on them, with a letter and number on each card. Information gain is scaled by the total information available (see Equation 2). For this sequential task, this was calculated at the level of the *type* of information that the participants were choosing between, and not at the level of individual tokens of these data types—that is, the cards. In the standard selection task, in which only four cards, one of each type, are presented all at the same time, this type-token ambiguity does not arise. However, when confronted with a sequential stream of potential information, where the number of tokens to be experienced and, so, the total information are unknown, it is possible to compute the relative information gains only at the type level.

7. According to Oberauer et al. (1999), the probabilities and positivity bias model should not apply when categories are *coherent*—that is, when all members share a common feature. All the categories used in this experiment were coherent. So, although fitting the data well, it should not apply to these results (or to Oaksford et al.'s, 1999). We suggest that the model works simply because it exploits the prediction of the rational information gain model that rare categories are more informative—that is, the information gain model explains why such a simple strategy is normally adaptive.

APPENDIX A
The Predicted Probabilities of Selecting Each Card and the Best-Fit Parameter Values
for Each Model in the Model Comparison Exercise

Model	Condition	Cards				Parameters		
		<i>p</i>	<i>not-p</i>	<i>q</i>	<i>not-q</i>	<i>P(p)</i>	<i>P(q)</i>	ϵ
Inf. Gain	LL	.70	.26	.69	.26	.35	.36	.17
	LH	.76	.25	.50	.38	.37	.45	.38
	HL	.55	.37	.67	.31	.48	.43	.84
	HH	.38	.57	.36	.60	.56	.57	.22
						<i>P_C</i>	<i>P_F</i>	<i>P_E</i>
MM	LL	.73	.25	.74	.34	.26	.34	.27
	LH	.77	.21	.53	.41	.48	.41	.23
	HL	.53	.31	.64	.49	.36	.49	.47
	HH	.48	.27	.55	.62	.45	.62	.52
						<i>CE_{TF}</i>	<i>CE_{FT}</i>	<i>W_{FF}</i>
Val. S.	LL	.74	.30	.72	.30	.74	.72	.41
	LH	.77	.25	.51	.33	.77	.51	.50
	HL	.52	.33	.63	.27	.52	.63	.53
	HH	.55	.50	.50	.55	.55	.50	1.0
						<i>P(p)</i>	<i>P(q)</i>	<i>P_b</i>
P & P	LL	.72	.27	.70	.28	.28	.30	.02
	LH	.75	.23	.55	.43	.25	.45	.02
	HL	.58	.40	.68	.30	.42	.32	.02
	HH	.41	.58	.36	.62	.59	.64	.02

Note—Inf. Gain, information gain model; MM, mental models/logic model; Val. S., validating search model; P & P, probabilities and positivity bias model; L, low; H, high.

APPENDIX B
The Predicted Probabilities of Selecting Each Card When the Fits Were as
Parameter Free as Possible for Each Model in the Model Comparison Exercise

Model	Condition	Cards			
		<i>p</i>	<i>not-p</i>	<i>q</i>	<i>not-q</i>
Inf. Gain	LL	.75	.24	.64	.28
	LH	.73	.27	.44	.45
	HL	.46	.39	.80	.23
	HH	.46	.48	.38	.58
P & P	LL	.66	.21	.63	.23
	LH	.57	.30	.39	.48
	HL	.32	.54	.64	.23
	HH	.32	.55	.29	.58

Note—Inf. Gain, information gain model; P & P, probabilities and positivity bias model; L, low; H, high.

(Manuscript received February 12, 2002;
revision accepted for publication October 2, 2002.)