Optimal Data Selection in the Reduced Array Selection Task (RAST)

Mike Oaksford
University of Wales, Cardiff

Nick Chater, Becki Grainger, and Joanne Larkin
University of Warwick

The predictions of M. Oaksford and N. Chater's (1994) optimal data selection (ODS) model for the reduced array selection task (RAST) were tested in 4 experiments. Participants tested a hypothesis, *if p then q,* by selecting cards showing *q* or *not q* instances. In Experiment 1, where selections were made from different sized stacks of *q* and *not q* cards, as $P(q)$ increased, *not q* card selections rose, and *q* card selections fell, as predicted. Experiment 2 controlled for the possibility that stack height influenced responses; these results were also consistent with ODS. Experiment 3, which controlled further for this possibility, replicated Experiment 1. Experiment 4 addressed a final issue concerning the medium $P(q)$ condition by concentrating on initial card selections; the results were again consistent with ODS. Although generally consistent with the ODS model, these experiments also suggest some interesting revisions.

The psychology of reasoning appears to show that on some tasks people do not reason according to the rules of logic (e.g., Evans, 1982, 1989; Johnson-Laird & Byrne, 1991; Wason & Johnson-Laird, 1972). Some authors have therefore concluded that humans may be irrational (Stich, 1985, 1990). However, others have suggested that apparently irrational behavior occurs because participants do not regard these as logical tasks (e.g., Rips, 1990; Thagard, 1988). Consequently, some normative theory other than logic may better describe what participants should do in these experimental situations. Recently, Oaksford and Chater (1994, 1996) have suggested that the normative theory of optimal data selection (Lindley, 1956) may explain results from Wason's (1966, 1968) selection task. This task has been widely interpreted as raising doubts about human rationality (Cohen, 1981; Manktelow & Over, 1993; Stich, 1985, 1990).

In the selection task, an experimenter presents participants with four cards, each with a number on one side and a letter on the other, and with a rule of the form *if p then q;* for example, *if there is a vowel on one side (p), then there is an*

even number 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). 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. However, as few as 4% of participants make this response, with 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, 1970b).

Oaksford and Chater (1994) provided a rational analysis (Anderson, 1990) of the selection task based on Bayesian optimal data selection (Fedorov, 1972; Lindley, 1956; Mackay, 1992). They argued that participants' behavior reflects a strategy of optimizing the expected amount of information gained by turning each card. According to this view, the selection task is not a logical reasoning task but a task of probabilistic optimal data selection in inductive hypothesis testing. Oaksford and Chater also generalized their optimal data selection (ODS) model to all the main experimental results on the selection task. Specifically, it accounts for the nonindependence of card selections (Pollard, 1985), the negations paradigm (e.g., Evans & Lynch, 1973), the therapy experiments (e.g., Wason, 1969), the reduced array selection task (RAST; Johnson-Laird & Wason, 1970a), and work on so-called fictional outcomes (Kirby, 1994). Oaksford and Chater also showed how a related maximum expected utility model accounts for deontic versions of the selection task (e.g., Cheng & Holyoak, 1985), where participants must reason about how one ought to behave, including perspective and rule-type manipulations (e.g., Cosmides, 1989; Gigerenzer & Hug, 1992) and the manipulation of probabilities and utilities (Kirby, 1994).[1]

The status of Oaksford and Chater's (1994) model is contentious (see, e.g., Almor & Sloman, 1996; Evans & Over, 1996; Laming, 1996; and for a reply, Oaksford & Chater, 1996). On the one hand, support for this model is

---

---

[1] Copies of the Mathematica program (Wolfram, 1991) in which the ODS model is implemented are available on request from Mike Oaksford preferably via electronic mail.

derived from Oaksford and Chater's (1995a) reinterpretation of Sperber, Cara, and Girotto's (1995) results and the recent results of Manktelow, Sutherland, and Over (1995). The latter confirmed the prediction of Oaksford and Chater's expected utility model of the deontic selection task, which is that both *not q* card and *p* card selections should decrease as the probability that the rule is being violated decreases (see Oaksford & Chater, 1994, Table 10). On the other hand, Evans and Over (1996) argued that data from Pollard and Evans's (1983) experiments falsify the ODS model. Moreover, Green (1995) and Platt and Griggs (1995) claimed to show that probabilistic manipulations fail to have the effects predicted by Oaksford and Chater (1994) in the abstract selection task. However, Oaksford and Chater (1996) argued that the ODS model may be consistent with Pollard and Evans's data. Further, they argued that the probabilistic manipulations used by Green and by Platt and Griggs (1995) may have been confounded by other manipulations designed to force a logical response. Consequently, how these data bear on Oaksford and Chater's model is unclear.

The purpose of the present experiments is to provide more direct tests of the predictions of Oaksford and Chater's (1994) model for the RAST (Girotto, 1988; Girotto, Blaye, & Farioli, 1989; Girotto, Light, & Colbourn, 1988; Johnson-Laird & Wason, 1970a; Light, Blaye, Gilly, & Girotto, 1989; Wason & Green, 1984). In the RAST, participants typically have access only to the *q* and *not q* cards (hence, *reduced array*). We first introduce the ODS framework. We then introduce the RAST and the predictions made by ODS.

Optimal Data Selection

In this section, we outline Oaksford and Chater's ODS model. The outline is condensed, and we refer the reader to Oaksford and Chater (1994) for the complete description of the model.[2]

In Wason's (1966, 1968) selection task, participants confront a problem that is analogous to the scientist's problem of which experiment to perform. Scientists have a hypothesis (the conditional rule) to assess, and they aim to perform experiments (turn cards) likely to provide data (i.e., what is on the reverse of the card) bearing on its truth or falsity. Oaksford and Chater's (1994) model is based on contemporary Bayesian accounts of scientific inference, which reject Popper's (1935/1959) falsificationist view that only potentially falsifying evidence should be sought (Earman, 1992; Horwich, 1982; Howson & Urbach, 1989). Bayesian accounts adopt an explicitly *subjective* as opposed to a *frequentist* approach to probability. According to the subjective interpretation, probabilities are degrees of belief (Keynes, 1921; Ramsey, 1931) rather than limiting frequencies (e.g., Mises, 1939). Oaksford and Chater's model is about how prior beliefs affect judgments about the most informative data to select.[3] In particular, people's prior beliefs about the probabilities of the antecedents, *p* and consequents, *q*, of rules, *if p then q*, play a central role. In the experiments we report below, we attempt to manipulate participants' subjective degrees of belief by manipulating the objective frequencies of cards in the selection task.

Oaksford and Chater (1994) suggested that hypothesis testers should choose experiments (select cards) that provide the greatest possible "expected information gain" in deciding between two hypotheses: (a) that the task rule *if p then q* is true, that is, that *p*s are invariably associated with *q*s (although *q*s are not invariably associated with *p*s) and (b) that the occurrences of *p*s and *q*s are independent. Participants' prior degree of belief in (b) is $P(M_I)$, and their prior degree of belief in (a) is $P(M_D)$, that is, $1 - P(M_I)$, where M_I refers to the contingency table representing independence (*I*) between *p* and *q*, and M_D refers to the contingency table representing a dependency (*D*) between *p* and *q* (Oaksford & Chater, 1994). For most purposes, Oaksford and Chater have assumed that these are equally likely, that is, $P(M_I) = .5$. For each hypothesis, Oaksford and Chater define probability models (M_I and M_D) that derive from participants' prior beliefs about the probabilities of *p* and of *q* in the task rule. They define information gain as the difference between the uncertainty before receiving some data and the uncertainty after receiving that data, where they measure uncertainty by using Shannon–Wiener information (Shannon & Weaver, 1949; Wiener, 1948). This is the same approach to optimal data selection proposed by Lindley (1956). Thus, Oaksford and Chater define the information gain of data *D* as follows:

information before receiving *D*:

$$I(H_i) = -\sum_{i=1}^{n} P(H_i) \log_2 P(H_i), \quad (1)$$

information after receiving *D*:

$$I(H_i|D) = -\sum_{i=1}^{n} P(H_i|D) \log_2 P(H_i|D), \quad (2)$$

information gain: $I_g = I(H_i) - I(H_i|D)$.   (3)

The posterior probabilities, $P(H_i|D)$, are calculated with Bayes's theorem. Thus, information gain is the difference between the information contained in the prior probability of a hypothesis (H_i) and the information contained in the posterior probability of that hypothesis given some data *D*.

When choosing which experiment to conduct (i.e., which card to turn), participants do not know what that data will be (i.e., what will be on the back of the card), so they cannot calculate actual information gain. However, they can compute *expected* information gain. Expected information gain

---

[2] Oaksford and Chater (1994) embodied the distinction between the abstract (or indicative) task and the deontic task, first noted by Manktelow and Over (1987), by proposing different models for each case. However, this distinction has recently been questioned by Almor and Sloman (1996; see also Oaksford & Chater, 1996).

[3] It is important to note that Oaksford and Chater's (1994) model is not about hypothesis testing per se. It is about the optimal design of experiments (see Fedorov, 1972); that is, the question of where the best place to look for evidence to test a hypothesis is.

is calculated with respect to all possible data outcomes (e.g., for the p card: q and not q) and both hypotheses.

Given the expected information gains, a decision has to be made about which cards to select. Oaksford and Chater (1994) incorporated two aspects of the decision process in their measure. First, they introduced a noise factor by adding .1 to the information gain for each card. This allows that people may occasionally see the not p card as informational. Second, card selection is a competitive matter. To reflect this, Oaksford and Chater scaled their information gain measure by the mean value for all four cards. Consequently, card choice is relative to the total expected information gain available and is not determined by the absolute $E(I_g)$ value alone. The higher the proportion of the total $E(I_g)$ a card possesses the more likely it is to be selected. Oaksford and Chater referred to this derived measure as *scaled expected information gain, $SE(I_g)$*.[4]

Oaksford and Chater (1994) calculated the $SE(I_g)$ for each card, assuming that the properties described in p and q are rare (i.e., that they have a low probability of occurrence). Take the rule *all ravens are black*, for example; the probability that any given bird is a raven is low, as is the probability that it is black. The *rarity assumption* seems to apply to the vast majority of everyday categories that are used to construct hypotheses about the world. Moreover, there is evidence that people adopt this assumption from the literature on other reasoning tasks (Anderson, 1990; Klayman & Ha, 1987; see Oaksford & Chater, 1994). Oaksford and Chater (1996) pointed out that further evidence for the rarity assumption comes from the normative literature on Bayesian epistemology (e.g., Horwich, 1982; Howson & Urbach, 1989). Making a rarity assumption resolves the ravens paradox of non-Bayesian confirmation theory (Goodman, 1954/1983), whereby nonblack, nonravens (e.g., a pair of white socks) must confirm the hypothesis that all ravens are black. Consequently, there are strong normative and empirical grounds for Oaksford and Chater's (1994) assumption that people's strategies for dealing with conditional rules are adapted to the case where rarity holds.

Adopting the rarity assumption, the order in $SE(I_g)$ is as follows:

$$SE[I_g(p)] > SE[I_g(q)]$$

$$> SE[I_g(not\ q)] > SE[I_g(not\ p)].\quad (4)$$

This corresponds to the observed frequency of card selections in Wason's (1966) task: $n(p) > n(q) > n(not\ q) > n(not\ p)$, where $n(x)$ denotes the number of cards of type x selected. This account thus explains the predominance of p and q card selections as a rational inductive strategy. This ordering holds only when $P(p)$ and $P(q)$ are both low. Oaksford and Chater (1994) noted that task manipulations that suggest that this condition does not hold—when $P(p)$ or $P(q)$ or both are high—lead to alternative orderings, predominantly that,

$$SE[I_g(p)] > SE[I_g(not\ q)]$$

$$> SE[I_g(q)] > SE[I_g(not\ p)].\quad (5)$$

This ordering is more consistent with Popperian falsificationism, which favors the p and *not q* instances. The effects of rarity and its violation permit us to explain the range of results we outlined above and make definite predictions in the RAST.

Implementation and Sensitivity

The ODS model provides a rational analysis (Anderson, 1990, 1994) of the selection task, which suggests that manipulating $P(p)$ or $P(q)$ should lead to predictable variations in the proportions of cards selected in the selection task. However, as Oaksford and Chater (1994) discussed, the level of detail at which the model can make predictions also depends on how the cognitive system *implements* this model (see also Anderson, 1990). The critical question is how sensitive can we expect people to be to changes in $P(p)$ or $P(q)$ and consequently to changes in the $SE(I_g)$ values?

At one extreme, the cognitive system may implement the rational analysis directly (i.e., it may perform all the computations specified by the model). If this were the case, then varying the parameters of the model should lead to card selections that directly mirror the resulting $SE(I_g)$ values. At the other extreme, the cognitive system may implement this analysis with a hardwired and cognitively impenetrable (Pylyshyn, 1984) heuristic that has evolved to deal with an environment where rarity is the norm. If this were the case, then although our model would explain why selecting the p and q cards is an adaptive rational strategy, we could not predict any performance variation in response to variation in the model's parameters.

To explain the data, Oaksford and Chater (1994) have already assumed that people are sensitive to manipulations of $P(p)$ and $P(q)$, and consequently we regard the second possibility as implausible. We also regard the first possibility as implausible because, as Oaksford and Chater have argued, the full Bayesian analysis it assumes is likely to prove computationally intractable when scaled up to real human reasoning (see also Chater. & Oaksford, 1990; Oaksford & Chater, 1991, 1992, 1993). In summary, the truth must lie somewhere between these two extremes of perfect sensitivity to changes in $P(p)$ or $P(q)$ and no sensitivity to such changes.

Sensitivity may depend on a variety of factors. If people compute and mentally represent something analogous to $SE(I_g)$ values, then noise, imperfect transduction, or both may lead to reduced sensitivity. Moreover, there is quite a broad region where $SE[I_g(not\ q)] \approx SE[I_g(q)]$ as $P(p)$ or $P(q)$ vary. Depending on the discriminability between the mental analogues of $SE[I_g(not\ q)]$ and $SE[I_g(q)]$, this could lead to quite a broad region of uncertainty about which card to

---

[4] Noise is introduced prior to scaling by the mean value. This is because we envisage the mental correlates of the $E(I_g)$ measures being produced by an inherently noisy cognitive system. This is not unfamiliar from signal detection theory (McNicol, 1972). Indeed, we view the decision problem participants confront as analogous to the signal detection situation.
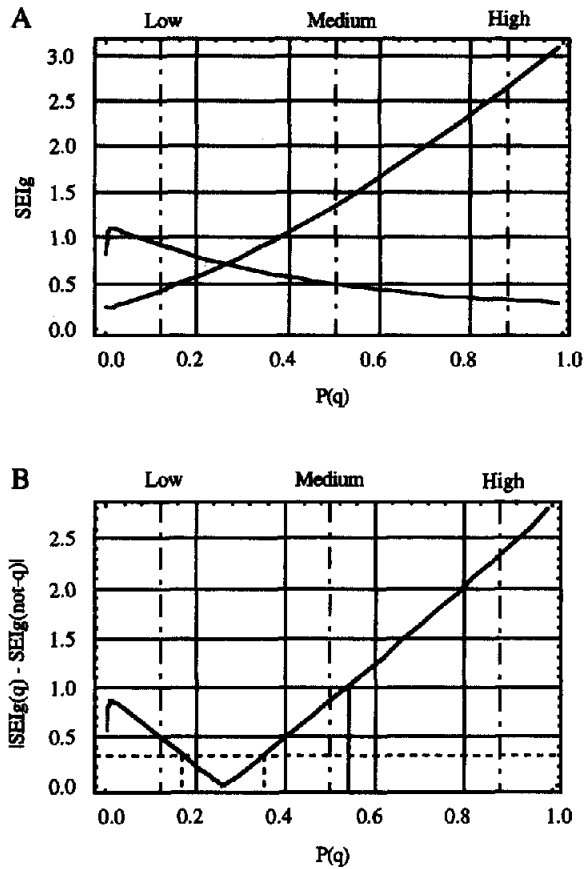
Figure 1. A: SEI_g plotted as a function of $P(q)$, with $P(M_I) = .5$ and $P(p) = P(q) - .001$. $SE[I_g(q)]$ is the decreasing function of $P(q)$ (over most of the range), and $SE[I_g(q)]$ is the increasing function. B: Discriminability, $|SE[I_g(not\ q)] - SE[I_g(q)]|$, plotted against $P(q)$. The vertical dashed lines labeled *Low, Medium*, and *High* correspond to the three values of $P(q)$ used in the subsequent experiments. The horizontal dashed line in (B) corresponds to a quite high level of sensitivity where little discrimination is required to determine which card to choose. SEI_g = scaled expected information gain; $P(q)$ = probability of q; $P(M_I)$ = participants' prior degree of belief that occurrences of ps and qs are independent.

choose. Figure 1 illustrates this point. The graph in figure 1A shows SEI_g plotted as a function of $P(q)$, with $P(M_I) = .5$ and $P(p) = P(q) - .001$; $P(p)$ was set so that we could observe the variation all the way down to the bottom end of the $P(q)$ scale.[5] $SE[I_g(q)]$ is the decreasing function of $P(q)$, and $SE[I_g(not\ q)]$ is the increasing function. The discriminability between the mental analogues of $SE[I_g(not\ q)]$ and $SE[I_g(q)]$ can be measured as the absolute difference between these two values, that is, as $|SE[I_g(not\ q)] - SE[I_g(q)]|$. This is plotted against $P(q)$ in Figure 1B. The level of discriminability required between the mental analogs of $SE[I_g(not\ q)]$ and $SE[I_g(q)]$ before which card to select is unambiguous will clearly vary. We have illustrated one possibility by the horizontal dashed line in Figure 1B. We have dropped vertical dashed lines to the $P(q)$ axis to illustrate the corresponding region of uncertainty. Between $P(q) = .18$ and $.35$ (approximately), participants may be relatively uncertain as to which card to turn. However,

according to Oaksford and Chater (1994) rarity is the norm. Consequently, participants may require higher discriminability to convince them that rarity has been violated than to convince them that they are in the default rarity situation. We have illustrated this possibility in Figure 1B by taking 1.0 as the higher level of discriminability required. The region of uncertainty now extends from $P(q) = .18$ to $.54$ (approximately). These values are clearly for illustrative purposes. Nonetheless, the sensitivity participants show toward variations in $P(q)$ clearly affects the predictions we can make in the selection task, and in deriving our predictions for the RAST we refer back to Figure 1.

Two further factors may affect sensitivity. First, sensitivity depends on how well people are calibrated to $P(p)$ or $P(q)$. People tend to overestimate the bottom end and underestimate the top end of the probability scale (Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978). Second, whether people use probabilistic information can depend on how it is presented. Gigerenzer, Hell, and Blank (1988) and Gigerenzer and Hoffrage (1995) have shown that people are more likely to use probability information if it is presented in a frequency format (e.g., 10 out of a 1,000, rather than in a probability format, i.e., .01). In the RAST, the probability information is always in a frequency format, which suggests that participants should use this information.

The Reduced Array Selection Task (RAST)

In a RAST, participants choose between the q and *not q* options only, hence *reduced array* (Johnson-Laird & Wason, 1970a; Wason & Green, 1984). The stimuli in the original RAST consisted of 30 colored shapes. The experimenter informs the participants that there are 15 black shapes and 15 white shapes, each of which is a triangle or a circle. The shapes are in two boxes, one containing the white shapes and the other containing the black shapes. On being presented with a test sentence (e.g., *All the triangles are black.*), participants have to assess the truth or falsity of the sentence by asking to see the *least* number of black or white shapes.[6] In Johnson-Laird and Wason's (1970a) research, although all participants chose some confirmatory black shapes (no participant chose more than 9), they all chose all 15 potentially falsificatory white shapes. Thus, where participants in effect perform multiple selection tasks, they tend to show falsificatory behavior. Wason and Green (1984) reported a variant on the RAST (see Oaksford & Chater, 1994), and Girotto and Light and their colleagues (Girotto,

1988; Girotto et al., 1988, 1989; Light et al., 1989) have used it in developmental studies with thematic content.

Oaksford and Chater (1994) suggested the following explanation for the basic findings of the RAST. The RAST makes explicit that the rule applies to a limited domain of cards or shapes that the experimenter describes as being in a box or in a bag (or in Wason & Green's, 1984, research as under the bar). The experimenter also informs participants that in this limited domain, there are equal numbers of q and *not q* instances. It follows that $P(q) = P(not\ q) = .5$, violating the rarity assumption. If participants are sensitive to these experimentally given frequencies, then this leads to a value of $SE[I_g(not\ q)]$ that is higher than $SE[I_g(q)]$. Consequently, ODS predicts more *not q* card selections than q card selections as is typically observed in the RAST.

In the experiments that follow, we tested this explanation of performance on the RAST by systematically varying $P(q)$. We did this by using stacks of cards rather than boxes of colored shapes. The number of cards in each stack was varied to achieve the probability manipulation. By varying these probabilities, we hoped to show that the proportions of q and *not q* cards selected would vary in accordance with the ODS model; that is, as $P(q)$ falls, q card selections rise, and *not q* card selections fall.

Experiment 1

ODS suggests that varying the ratio of q and *not q* cards—that is, varying $P(q)$—in the RAST will vary $P(p)$. This is because for the rules in the selection task to make sense $P(p) < P(q)$. This is also a constraint on Oaksford and Chater's (1994) model.[7] If this were not the case, then one could state nonsensical, obviously false rules (e.g., *all black things are ravens*). You know this is false without examining any data because you know that there are many black things other than ravens. For our predictions, this means that as $P(q)$ decreases, participants should decrease their subjective estimates of $P(p)$ by at least as much. (This factor was responsible for our setting $P(p)$ to $P(q) - .001$ in deriving Figure 1.)

In these experiments, we use three different values of $P(q)$, low, medium, and high (see Figure 1), and test whether this variation leads to changes in q and *not q* card selections in the RAST. We use the computed SEI_g values only as a guide to predicting the frequency of card selections because of the uncertainty about the level of sensitivity participants are likely to display to variations in $P(q)$. These possible variations suggest two broad levels at which predicting the pattern of card selection frequencies seems feasible. First, we consider the predictions that can be made if participants are relatively insensitive to variations in $P(q)$. We call these the Level 1 predictions. In deriving these predictions, we refer to Figure 1. Following Oaksford and Chater's (1994) research, we assume that card selection frequencies are monotonically related to SEI_g values. According to this assumption, most of our predictions can be seen to follow directly from Figures 1A and 1B. At Level 1, three such predictions of increasing levels of detail can be derived.

Level 1 Predictions

1. Even if participants are relatively insensitive to variation in $P(q)$, the model predicts that the difference in the frequency of q card and *not q* card selections, that is, $n(not\ q) - n(q)$, should mirror the relationship between $SE[I_g(not\ q)] - SE[I_g(q)]$ and $P(q)$. This relationship is partly illustrated in Figure 1B. However, Figure 1B shows the absolute value of $SE[I_g(not\ q)] - SE[I_g(q)]$ so at the point where $|SE[I_g(not\ q)] - SE[I_g(q)]| = 0$, the function should be viewed as continuing to be negative as $P(q)$ continues to decrease. Consequently, the model predicts that there should be a significant trend such that $n(not\ q) - n(q)$ increases as $P(q)$ increases.

2. Moreover, in principle, the model predicts that there should be significant pairwise differences between all conditions on the $n(not\ q) - n(q)$ measure. However, because lack of sensitivity may affect both the low and medium $P(q)$ conditions, in practice there may be no significant difference between these two conditions (see Figure 1B).

3. Referring to Figure 1A, the model predicts individual trends such that as $P(q)$ increases, the frequency of q card selections, $n(q)$, should fall, and the frequency of *not q* card selections, $n(not\ q)$, should rise.

As we discuss further in the General Discussion, it is important to note that no other theory of RAST performance makes similar predictions even at this level of detail. For example, although Kirby (1994) predicted similar effects on $n(not\ q)$ as a result of variation in $P(p)$, his account could not predict similar effects on $n(q)$ when varying either $P(p)$ or $P(q)$. Consequently, if these predictions are confirmed at any level of detail, then these data will raise serious questions for other theories of reasoning.

Because insensitivity may affect both the low and medium $P(q)$ conditions, we cannot unequivocally predict differences between the frequencies of q and *not q* card selections *within* conditions, except for the high $P(q)$ condition. However, if participants are relatively sensitive to variations in $P(q)$, then the following simple effects can also be predicted.

Level 2 Predictions

1. In Figure 1A, in the low $P(q)$ condition there should be more q card selections than *not q* card selections. In the medium and high $P(q)$ conditions, there should be more *not q* card selections than q card selections.

2. In principle, the model also predicts the following interactions (differences between the above simple effects). There should be more q card selections and fewer *not q* card selections in the low versus the high $P(q)$ conditions, in the medium versus high $P(q)$ conditions, and in the medium

---

[7] The actual constraint on the model is that $P(q) \geq P(p)P(M_D)$. However, in the RAST, participants rapidly come to believe that the task rule is true, because all the cards conform to the rule, which means that $P(M_D) \approx 1$. Moreover, this constraint arises in the computation of *expected* values of $P(p)$ and $P(q)$ with respect to the two models. Actual values of $P(p)$ and $P(q)$ must respect the $P(q) \geq P(p)$ constraint for the reasons given in the text.

versus low $P(q)$ conditions. The last two interactions should be full crossover interactions. However, because lack of sensitivity may affect both the low and medium $P(q)$ conditions, again, in practice there may be no significant interaction between these two conditions (see Figure 1).

These predictions also rely on the assumption that the RAST procedure does not substantially alter participants' assessments of the ODS model's parameters. The RAST may violate this assumption because it is a sequential sampling paradigm in which the evidence is revealed at each trial. Hence, participants may continuously update the parameters of the model. However, this seems unlikely to affect our model's predictions. First, all the data are consistent with the dependence model. Consequently, Bayesian revision of $P(M_I)$ can lead only to lower values; that is, $P(M_D)$ can only get higher. But as Oaksford and Chater (1994) observed, the predictions of the ODS model are relatively insensitive to variation in $P(M_I)$. Nonetheless, as $P(M_D)$ tends to 1, so the $SE(I_g)$ values tend to 1. (This is only because of the scaling factor; the $E(I_g)$ values tend to 0.) This means that as the evidence convinces you that the rule is true, which evidence is relevant becomes harder to discriminate. This could lead to some unpredicted switches of attention.

Second, another way to affect the model's predictions is for sampling to affect participants' estimates of $P(p)$ and $P(q)$. However, participants know $P(q)$ from the outset because the experimenter tells them the numbers of q and $not\ q$ cards that are in the stacks from which they start to sample cards. Consequently, information gained from sequential sampling should not alter the value of this parameter. Further, if $P(q)$ remains constant, then under the constraint that $P(p) \le P(q)$, the predictions of the ODS model remain the same, as we saw above.

Method

Participants. There were 36 participants. All participants were recruited from the Department of Psychology, University of Wales, Bangor participant panel. Each participant was paid £2.50 for his or her participation. None of the participants had any prior knowledge of Wason's (1966) selection task.

Design. Participants were assigned randomly to one of three conditions, so that there were 12 participants in each condition: a high probability condition, $P(q) = .83$ (⅚); a medium probability condition, $P(q) = .50$ (½); and a low probability condition, $P(q) = .17$ (⅙).

Materials. The materials consisted of two packs of 25 cards each. One pack depicted red circles on one side, and the other pack depicted blue triangles on one side. The obverse side of all the cards in each pack was blank. The rule used in this experiment was *all the triangles (p) are blue (q)*. Two stacks of cards were placed before a participant. One stack was drawn from the pack of red circles and the other from the pack of blue triangles. Both stacks were placed before participants with the blank faces uppermost so that they could not see the colored shapes on the cards. Each stack had a label behind it. The stack of red circles had a label reading *Red Shapes (not q* cards), and the stack of blue triangles had a label reading *Blue Shapes (q* cards). In the high probability condition, there were 25 cards in the q card stack and 5 cards in the *not q* card stack. In the medium probability condition, there were 15 cards in

the q card stack and 15 cards in the *not q* card stack. In the low probability condition, there were 5 cards in the q card stack and 25 cards in the *not q* card stack.

In the RAST, a "prompt" card is usually included among the q cards. This card is an instance of a *not p* and q card (i.e., in this experiment, a blue circle). The purpose of the prompt card is to indicate the irrelevance of the q cards; "such an item is obviously irrelevant to the claim that all triangles are [blue] and hence it follows that the [blue] shapes provide no information" (Wason & Green, 1984, p. 600). According to the logic of ODS, this argument is not valid. Although finding blue circles may be uninformative concerning the truth or falsity of *all the triangles are blue*, this does not mean that finding blue *triangles* is similarly uninformative. From the perspective of ODS, therefore, the prompt introduces an unwanted confound into the RAST procedure (see also Wason & Green, 1984, p. 600). In this experiment, the prompt card was therefore omitted from the q card stack.

Procedure. Participants were tested individually. On entering the experimental room, participants were seated in front of a table where the experimental materials were laid out. The experimenter then read out the following instructions (based on Johnson-Laird & Wason's, 1970a, instructions):

> Before you are two stacks of cards. Each card in these stacks has a blue shape or a red shape on it. The shapes on the cards are triangles or circles.
> Your task is to prove whether the following claim is true or false of these cards:
> ALL THE TRIANGLES ARE BLUE.
> You ask me to turn over a card from either stack (blue shapes or red shapes), then, when you have examined it, you ask me to turn another card—again from either stack. We continue in this way until you are satisfied that the rule is true or false. There is no hurry, but I want you to do the task economically, i.e., to examine as few cards as you need.
> Some information that may help you decide which cards to choose is that there are. . .

The information participants were then given depended on the condition. In the high probability condition, they were told there were 25 blue shapes and 5 red shapes, in the medium probability condition they were told there were 15 blue shapes and 15 red shapes, and in the low probability condition they were told there were 5 blue shapes and 25 red shapes. Participants were then told, "When you have reached a decision about whether the claim is true or false, and hence you do not need to turn any more cards, please inform the experimenter."

In Johnson-Laird and Wason (1970a), "if at any point it was claimed that the sentence had been proved true, when it had not been proved, the participant was told he was wrong and asked to continue with the task" (p. 51). We used the same instruction, but with one change in how it was applied. This instruction can be interpreted only to mean that if participants had not looked at all the *not q* cards and determined that none was also a *p* card, then they were told they were wrong. This criterion for terminating the experiment is based on the strategy of falsification. However, according to ODS, in the present experiment, different stacks were relevant in different conditions, so in the low probability condition, the q cards were most informative. Therefore, we predicted that participants should select all of these cards and only some of the *not q* cards. Conversely, in the high probability condition the *not q* cards were the most informative; therefore, we predicted that participants should select all of these cards and only some of the q cards. In all conditions, we therefore terminated the experiment when participants announced that the rule was either true or false and one stack was exhausted; that is, we laid down no prescriptions about which stack had to be exhausted before we allowed a

participant to terminate the experiment. This termination condition was used in all the experiments reported in this article. In practice, although the possibility of falsity is mentioned in the instructions, because the rule was always true of the cards in the stacks, participants always announced that the rule was true. (No participants gave the aberrant response that the rule was false.)

Results and Discussion

We first converted each participant's data into proportions of cards selected because of the unequal numbers of q and not q cards involved in each condition.[8] We show the proportions of each card selected in Table 1. We then transformed the data by using the arcsine transform so that the data would meet the underlying assumptions of the analysis of variance (ANOVA; Snodgrass, 1977, pp. 379–382). To test Level 1, Prediction 1 and Level 1, Prediction 2 of ODS, we linearly combined these measures (Lindeman, Merenda, & Gold, 1979) by taking the proportion of q card selections (arcsine corrected) from the proportion of not q card selections (arcsine corrected) to produce a composite variable that equates with discriminability. Oaksford and Stenning (1992) used a similar measure that they called the consequent falsification index (CFI). We show the values of CFI by $P(q)$ (low, medium, and high) in Table 1.

At Level 1, ODS predicts that CFI should increase as $P(q)$ increases. We therefore performed a linear contrast with $P(q)$ (low, medium, and high) as a between-subjects variable and with CFI as the dependent variable. As predicted, there was a highly significant linear contrast, $F(1, 33) = 26.68$, $MSE = 3.22$, $p < .0001$, such that CFI increased with $P(q)$. This confirmed Prediction 1, Level 1 of ODS. At Level 1, ODS also predicts that CFI should be higher in the high $P(q)$ condition than in the other two conditions and that CFI should be higher in the medium $P(q)$ condition than in the low $P(q)$ condition. Post hoc Newman–Keuls tests confirmed all these predictions at the .05 level. These tests confirmed Prediction 2, Level 1 of ODS in this experiment.

At Level 1, ODS further predicts individual trends in the proportions of q and not q cards selected as $P(q)$ varies (Prediction 3, Level 1). To test these predictions, we used similar between-subjects ANOVAs with the arcsine transform of the proportion of cards selected as the dependent variable. Consistent with the Level 1 predictions of ODS, the trend for a lower proportion of q card selections as $P(q)$ increases was highly significant, $F(1, 33) = 15.88$, $MSE =$

Table 1
Mean Proportion of q Cards, Mean Proportion of not q Cards, and Consequent Falsification Index (CFI; Arcsine Corrected) for Each Level of P(q) in Experiment 1

| $P(q)$ | Mean Prop. q | | Mean Prop. not q | | CFI | |
	M	SD	M	SD	M	SD
Low ($\frac{1}{6}$)	.88	.30	.35	.35	−1.51	1.86
Medium ($\frac{1}{2}$)	.68	.47	.68	.34	−0.02	2.27
High ($\frac{5}{6}$)	.25	.36	1.00	.00	2.28	1.03

Note. Prop. = proportion; $P(q)$ = probability of q.

1.36, $p < .0001$, as was the trend for a higher proportion of not q card selections as $P(q)$ increases, $F(1, 33) = 32.86$, $MSE = 0.65$, $p < .0001$. These simultaneous effects on both the q and the not q cards are predicted only by the ODS model.

The Level 2 predictions of ODS rely on people being quite sensitive to variation in $P(q)$ and consequently to differences in $SE[I_g(q)]$ and $SE[I_g(\text{not } q)]$ within conditions. At this level, ODS predicts that in the low $P(q)$ condition, there should be more q card selections and fewer not q card selections and that in the medium and high $P(q)$ conditions, there should be more not q card selections and fewer q card selections. Qualitatively, these predictions are supported by the CFIs in each condition except for the medium condition. On average, CFI was negative for the low and medium $P(q)$ conditions, albeit only marginally for the medium $P(q)$ condition, indicating more q than not q card selections. In contrast, on average, CFI was positive in the high $P(q)$ condition, indicating more not q than q card selections.

We further investigated these results by conducting a 3 × 2 mixed ANOVA with $P(q)$ as the between-subjects variable and card as the within-subjects variable and with proportions of cards selected (arcsine corrected) as the dependent variable. We used simple effects comparisons at each level of $P(q)$ to test the above predictions. At the low $P(q)$ level, the proportion of q cards selected was significantly higher than the proportion of not q cards selected, $F(1, 33) = 8.50$, $MSE = 1.61$, $p < .01$. Moreover, at the high $P(q)$ level, the proportion of not q cards selected was significantly higher than the proportion of q cards selected, $F(1, 33) = 19.27$, $MSE = 1.61$, $p < .0001$. However, at the medium $P(q)$ level, the proportion of not q cards selected was not significantly higher than the proportion of q cards selected, $F(1, 33) < 1$. Apart from the final comparison, these results confirmed Level 2, Prediction 1 of ODS.

Finally at Level 2, ODS predicts the following interactions between these simple effects: There will be more q card selections and fewer not q card selections in the low $P(q)$ condition than in the medium and the high $P(q)$ conditions and more q card selections and fewer not q card selections in the medium $P(q)$ condition than in the high $P(q)$ condition (Prediction 2, Level 2). To test these pairwise interactions between conditions, we treated q and not q cards as levels of a single within-subjects cards variable and performed three separate 2 × 2 mixed ANOVAs with proportion of cards selected (arcsine corrected) as the dependent variable. For the low versus high $P(q)$ comparison, there was a significant crossover interaction, $F(1, 22) = 38.12$, $MSE = 1.13$, $p < .0001$, such that there was a higher proportion of q cards selected and a lower proportion of not q cards selected in the

[8] Proportions were used because of the termination conditions of the experiment that one stack has to be exhausted. By Bayesian revision, this may mean that participants were convinced that the rule was true before they were allowed to stop sampling. In these circumstances, we assume that participants continued to sample from the stack suggested by ODS when they became convinced of the rule's truth. As we point out later on, this assumption may need to be revised in the light of the data.

low $P(q)$ condition than in the high $P(q)$ condition. There was also a significant interaction for the medium versus high $P(q)$ comparison, $F(1, 22) = 10.12$, $MSE = 1.55$, $p < .005$, such that there was a higher proportion of q cards selected and a lower proportion of *not* q cards selected in the medium $P(q)$ condition than in the high $P(q)$ condition. However, the interaction between the medium and the low $P(q)$ conditions failed to reach significance at the 5% level, $F(1, 22) = 3.12$, $MSE = 2.15$, $p = .091$. Apart from the final interaction, these results confirmed Prediction 2, Level 2 of ODS.

These findings are consistent with the predictions of ODS, which predicts that as $P(q)$ and hence $P(p)$ rises, q card selections should decrease and *not* q card selections should increase. In particular, Experiment 1 demonstrates that in a task where the *not* q card selection normally dominates, a probabilistic manipulation can lead to more q card selections. Other than the failure to observe the predicted results for the medium $P(q)$ condition, all the Level 1 and the Level 2 predictions of the ODS model were confirmed in this experiment. The confirmation of the Level 2 predictions seems consistent with participants being quite sensitive to variations in $P(q)$.

In part, these results are also consistent with the standard RAST results. In the medium $P(p)$ condition, $P(p) = P(q) = .5$, as in the standard RAST. The proportion of *not* q cards selected was 68%, which is far higher than the 4% to 12% found in the standard selection task. This is the standard RAST result. However, in Experiment 1 the proportion of q cards selected was the same as the proportion of *not* q cards selected, whereas in the standard RAST it is typically lower. Moreover, the number of participants selecting all the q and only some of the *not* q cards was higher than the number of participants selecting all the *not* q and only some of the q cards, which does not replicate the standard RAST. These results are not consistent with ODS, which predicts more *not* q card than q card selections at these values of $P(p)$ and $P(q)$.

One possible reason for the discrepancy is that we did not use a prompt card to try to cue participants that the q cards were irrelevant. Another possible reason is that, as we discussed when introducing this experiment, participants were not perfectly sensitive to variations in $P(q)$. As we mentioned, because rarity is the norm, participants may require quite high discriminability to prefer *not* q to q, which equates with higher values of $P(q)$ (see Figure 1B). This situation could produce the pattern of results where in the mid range of $P(q)$ values, participants are uncertain of which card to choose. We return to this issue later, specifically in Experiment 4. We first discuss other possible explanations for the results of Experiment 1. We address these possible explanations in the next two experiments.

In the low $P(q)$ condition, the smallest stack was the q stack. In the high $P(q)$ condition, the smallest stack was the *not* q stack. In these circumstances, there are other possible explanations of Experiment 1 that we investigate in the next two experiments. First, these results may be found if participants were selecting randomly from the stacks and terminating when one stack was exhausted. In these circumstances, the proportion of cards selected from the small stack

would always be higher and would always correspond to the stack that ODS predicts should yield the highest proportion of selections. Second, participants could have been choosing cards from the stack containing the fewest cards; once they exhausted this stack, they announced that the rule was true.

Such a "small stack bias" or random selection does not account for why participants also selected cards from the stack containing the most cards. It could be argued that a few logically competent falsifiers (selecting *not* q cards) in the low condition and a few committed confirmers (selecting q cards) in the high condition could explain this finding. However, even with this modification, neither small stack bias nor random selection can explain these data. First, small stack bias must predict that *most* participants should not pick any cards from the stack containing the most cards; but in the low $P(q)$ condition, only 1 participant did not pick one or more cards from the stack containing the most cards, and in the high $P(q)$ condition, only 4 participants did not pick one or more cards from the stack containing the most cards. Second, random selection could not predict that participants would select more cards from the stack containing the most cards. However, on average, in the low $P(q)$ condition, participants selected 8.67 cards from the stack containing the most cards, and in the high $P(q)$ condition, they selected 6.25 cards from the stack containing the most cards. These results are not consistent with participants selecting cards at random and terminating when one stack was exhausted.

Another possibility is that participants prefer *not* q cards in the low $P(q)$ condition in some fixed ratio to the q cards. In the low $P(q)$ condition, although participants chose all the q cards and only some of the *not* q cards, they chose approximately twice as many *not* q as q cards, giving a 2:1 preference for *not* q cards. If participants had selected in this ratio between stacks, they could still have exhausted the small stack while preferring *not* q cards in this condition.[9] However, according to this proposal, there is no reason to predict that people's preferences should change between conditions. Consequently, participants should prefer *not* q to q cards in the fixed ratio 2:1; that is, they should also have selected twice as many *not* q as q cards in the medium and high conditions. However, in the medium condition this ratio was 1.0, and in the high condition it was .8. These values are not consistent with the proposal that participants are selecting cards in a fixed ratio.

An issue raised by this proposal concerns our use of the proportion of cards selected as the dependent variable in these analyses. This was the appropriate measure because to

---

[9] It is legitimate to question what sense can be made of "preference" in this context such that not all of the "preferred" option was selected. In this task, both options q and *not* q were equally accessible, and so it would be expected that the most preferred option would be exhausted. Moreover, the only way that it seems to make sense to talk of preferring one type of evidence to another in some fixed ratio (as opposed to, say, one kind of fruit to another) is in terms of the very probabilistic analysis that the ODS model offers. For example, the values of $P(q)$—and consequently of $P(p)$—may suggest that one is twice as likely to find some discriminatory evidence on the *not* q than on the q cards.

achieve the probability manipulation the stack sizes were unequal, which meant that participants had unequal opportunities to select cards from the different stacks. Using the proportion allows us to control for this unequal number of opportunities to select cards. However, this leaves open the possible confounds we have discussed. Although, as we have argued, Experiment 1 itself provides evidence that these results are not confounded in the ways that we have mentioned, in Experiments 2–4, to which we now turn, we further controlled for these possibilities.

Experiment 2

If altering the number of cards available induces participants to choose cards at random, from the stack containing the fewest cards, or in a fixed ratio, then participants should do the same if we use the p and $not\ p$ cards in the RAST. ODS predicts that the p card should be preferred to the $not\ p$ card at all levels of $P(p)$. These other possibilities, in contrast, predict that the $not\ p$ card should be preferred to the p card in the high $P(p)$ condition. Using the antecedent cards in a RAST procedure would also allow us to test the predictions of ODS for these cards. ODS also predicts that as $P(p)$ increases, p card selections should decrease and $not\ p$ card selections should increase. However, the termination criterion for the RAST implies that all cards in the most preferred stack will be selected. ODS must therefore predict that the p stack will always be exhausted because it predicts that the p cards should be preferred in all conditions. Consequently, because of the termination conditions for the RAST, p card selections are predicted to always be at ceiling, and hence the predicted trend for this card is unlikely to be observed.

Method

Participants. There were 36 participants. All participants were recruited from the Department of Psychology, University of Wales, Bangor participant panel. Each participant was paid £2.50 for his or her participation. None of the participants had any prior knowledge of Wason's (1966) selection task.

Design. Participants were assigned randomly to one of three conditions, so that there were 12 participants in each condition: a high probability condition, $P(p) = .83$ (⅚); a medium probability condition, $P(p) = .50$ (½); and a low probability condition, $P(p) = .17$ (⅙).

Materials. The materials were the same as those used in Experiment 1. The same rule—*all the triangles (p) are blue (q)*—was also used. However, because the choice of cards was now between the p and the *not p* cards, the stack of red circles had a label reading *Circles (not p* cards), and the stack of blue triangles had a label reading *Triangles (p* cards). The number of p and *not p* cards in each condition directly mirrored the number of q and *not q* cards, respectively, in each condition in Experiment 1.

Procedure. The procedure was the same as in Experiment 1 except for the following changes to the instructions to reflect that the choice of cards was now between the p and *not p* cards. Sentences 2 and 3 were replaced with "Each card in these stacks has a coloured triangle or a circle on it. The colours on the cards are blue or red," and the description in parentheses at the end of the fifth sentence was replaced with "triangles or circles." The information participants were then given, depending on the condi-

Table 2
Mean Proportion of p Cards and Mean Proportion of not p Cards for Each Level of P(p) in Experiment 2

	Mean prop. p		Mean prop. not p	
P(p)	M	SD	M	SD
Low (⅙)	.92	.29	.14	.30
Medium (½)	1.00	.00	.19	.31
High (⅚)	.93	.25	.42	.42

Note. Prop. = proportion; $P(p)$ = probability of p.

tion, was the same as in Experiment 1 except for the substitution of *triangles* for *blue* and *circles* for *red*.

Results and Discussion

Our goal in this experiment was to control for the possibility that participants either were selecting cards at random or were biased toward the smaller stack. This predicts that participants should opt for the smallest stack in the low and high $P(p)$ conditions. Consequently, they should select more p cards than $not\ p$ cards in the low $P(p)$ condition but more $not\ p$ cards than p cards in the high $P(p)$ condition. In contrast, ODS predicts that participants should make more p than $not\ p$ card selections regardless of $P(p)$. To test these competing hypotheses, we first transformed the raw data in the same way as in Experiment 1. We then performed a 2 × 3 mixed ANOVA with card type (p and $not\ p$) as a within-subjects variable and $P(p)$ (low, medium, and high) as a between-subjects variable, with the arcsine of the proportion of cards selected as the dependent variable. There was a significant main effect for card type, $F(1, 33) = 69.40$, $MSE = 1.18$, $p < .0001$, but not for $P(p)$, $F(2, 33) = 2.92$, $MSE = 0.42$, $p > .10$, with no significant interaction, $F(2, 33) = 1.16$, $MSE = 1.18$, $p > .30$. We show proportion of cards selected in each condition in Table 2. Table 2 shows that consistent with ODS, the proportion of p card selections is higher than the proportion of $not\ p$ cards selected for all levels of $P(p)$. Simple main effects comparisons revealed that the proportion of p cards selected was significantly higher than the proportion of $not\ p$ cards selected in the low $P(p)$, $F(1, 33) = 28.51$, $MSE = 1.18$, $p < .0001$, in the medium $P(p)$, $F(1, 33) = 30.47$, $MSE = 1.18$, $p < .0001$, and crucially in the high $P(p)$, $F(1, 33) = 12.76$, $MSE = 1.18$, $p < .001$, conditions. These results, although consistent with ODS, are not consistent with random selection or with a small stack bias. In particular, both possibilities predict that $not\ p$ should be preferred to p in the high $P(p)$ condition, which it was not. It is therefore unlikely that participants adopted a small stack bias or selected cards at random in Experiment 1.

We further investigated any trends for the p and $not\ p$ card selection frequencies. Consistent with ODS, there was a trend for a higher proportion of $not\ p$ card selections as $P(p)$ increased that just failed to reach significance at the 5% level, $F(1, 33) = 3.67$, $MSE = 1.16$, $p = .064$. As predicted, we did not observe a significant trend for the p card, $F(1, 33) < 1$.

These results indicate that altering the number of cards available does not induce participants to select cards at random or from the stack containing the fewest cards. Participants consistently selected more p cards than *not p* cards for all levels of $P(p)$. This means that any attempt to explain the results of Experiment 1 as an artifact of altering the number of cards available must provide a different account of Experiment 2, an almost identical task.

The trend for more *not p* card selections as $P(p)$ increases is consistent with ODS. As predicted, p card selections were very near ceiling at all levels of $P(p)$. However, the p card's dominance may call into question whether Experiment 2 successfully addressed small stack bias. Therefore, we further investigated this bias in Experiment 3.

Experiment 3

The dominance of the p card may call into question Experiment 2's ability to act as a control for a small stack bias. Every theory of the selection task predicts that the p card should predominate.[10] Although the dominance of the p card overrides the small stack bias for the antecedent cards in Experiment 2, this does not mean that this bias or a random selection of cards was not the cause of the behavior observed for the consequent cards in Experiment 1. In Experiment 3, we controlled for this potential confound by replicating Experiment 1 but keeping the stack sizes the same. We achieved this by having the experimenter deal 10 cards from different sized packs. Participants then had to select from the two now equally sized stacks.

Method

Participants. There were 36 participants. All participants were recruited from the Department of Psychology, University of Warwick participant panel. Each participant was paid £2.00 for his or her participation. None of the participants had any prior knowledge of Wason's (1966) selection task.

Design. Participants were assigned randomly to one of three conditions, so that there were 12 participants in each condition: a high probability condition, $P(q) = .83$ (⅚); a medium probability condition, $P(q) = .50$ (½); and a low probability condition, $P(q) = .17$ (⅙).

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 obverse side of all the cards in each pack was uniformly patterned to prevent show-through of the colored shapes. The rule used in this experiment was *all the triangles (p) are blue (q)*. In this experiment, the stacks from which participants selected cards were drawn from two larger packs that were placed in front of participants. One pack consisted of red circles and the other of blue triangles. Both packs were placed before participants with the patterned faces uppermost so that they could not see the colored shapes on the cards. Each pack had a label behind it. The pack of red circles had a label reading *Red Shapes* (*not q* cards), and the pack of blue triangles had a label reading *Blue Shapes* (*q* cards). In the high probability condition, there were 100 cards in the *q* card pack and 20 cards in the *not q* card pack. In the medium probability condition, there were 60 cards in the *q* card pack and 60 cards in the *not q* card pack. In the low probability condition, there were 20 cards in the *q* card pack and 100 cards in the *not q* card pack.

Procedure. The procedure was the same as in Experiment 1, with the following changes to the instructions.

> Before you are two packs of cards. One pack contains cards with red shapes on one side and the other pack contains cards with blue shapes on one side. The shapes on the cards are triangles or circles. I will deal you 10 cards with a red shape on one side and 10 cards with a blue shape on one side.
>
> Your task is to prove whether the following claim is true or false of these cards:
>
> ALL THE TRIANGLES ARE BLUE.
>
> You ask me to turn over a card from either stack (blue shapes or red shapes), then, when you have examined it, you ask me to turn another card—again from either stack. We continue in this way until you are satisfied that the rule is true or false. There is no hurry, but I want you to examine as few cards as you need.
>
> Some information that may help you decide which cards to choose is that, there are ...

The information participants were then given depended on the condition. In the high probability condition, they were told that the packs from which the stacks of cards were drawn contained 100 blue shapes and 20 red shapes, in the medium probability condition, 60 blue shapes and 60 red shapes, and in the low probability condition, 20 blue shapes and 100 red shapes.

It is possible that these instructions created an ambiguity about whether the rule applied (a) to just the 20 cards in the two stacks or (b) to all the cards in the pack. Only the latter interpretation is consistent with altering participants' beliefs about $P(q)$ in the intended directions. The first interpretation (a) would leave $P(q) = .5$ in all conditions. There are two points to make here. First, the additional information participants were given, which they are told "may help ... [to] ... decide which cards to choose," was inconsistent with this interpretation (a) according to which, information about the distribution of colors in the *packs* must be irrelevant. Therefore, it is unlikely that having been told that information about the packs was relevant to their card choice, participants should then have interpreted the rule as applying only to the cards in the two stacks. Second, according to ODS, if $P(q) = .5$ in all conditions, then there should have been no differences in card selections between conditions. Consequently, interpretation (a) can be dismissed to the same extent that our experimental hypotheses were confirmed.

One final concern with the revised procedure in Experiment 3 is that with the intended interpretation (b), participants can never be certain that the rule is actually true because it applies to cards that they are not given the opportunity to examine. Nonetheless, given the random sample of 20 cards drawn from this population, they can be highly confident that it does apply. Moreover, any lack of certainty must predict that participants would never declare the rule true on the basis of sampling the cards in the stacks. Further, it would predict that participants would exhaust both stacks without seeking to terminate the experiment. To anticipate our findings, neither of these predictions were confirmed. Consequently, it would appear that interpretation (b) does not significantly alter the way participants viewed the task.

---

[10] It is reasonable, however, to question the extent to which other theories predict the dominance of the p card. Most other theories provide informal accounts that incorporate this dominance post hoc. In contrast, Oaksford and Chater's (1994) ODS model applied ideas from optimal data selection provided by Lindley (1956) to the selection task from which the p card dominance emerges as a consequence.

Results and Discussion

We analyzed the results of Experiment 3 in the same way as we analyzed those in Experiment 1. At Level 1, ODS predicts that CFI should increase as $P(q)$ increases (see Table 3). We therefore performed a linear contrast with $P(q)$ (low, medium, and high) as a between-subjects variable and with CFI as the dependent variable. As predicted, there was a significant linear contrast, $F(1, 33) = 4.53$, $MSE = 4.56$, $p < .05$, such that CFI increased with $P(q)$. This analysis confirmed Prediction 1, Level 1 of ODS. At Level 1, ODS also predicts that CFI should be higher in the high $P(q)$ condition than in the other two conditions and that CFI should be higher in the medium $P(q)$ condition than in the low $P(q)$ condition. Post hoc Newman–Keuls tests confirmed the first two predictions at the .05 level. However, there was no significant difference between the medium $P(q)$ and the low $P(q)$ conditions. These tests confirmed all but one of the Prediction 2, Level 1 predictions of ODS in this experiment.

At Level 1, ODS further predicts individual trends in the proportions of q and not q cards selected as $P(q)$ varies (Prediction 3, Level 1). To test these predictions, we used similar between-subjects ANOVAs with the arcsine transform of the proportion of cards selected as the dependent variable. Consistent with the Level 1 predictions of ODS, the trend for a lower proportion of q card selections as $P(q)$ increases was significant, $F(1, 33) = 6.28, MSE = 1.53, p < .025$. However, the trend for a higher proportion of not q card selections as $P(q)$ increased failed to reach significance at the 5% level, $F(1, 33) = 2.78, MSE = 1.12, p = .10$.

The Level 2 predictions of ODS rely on participants being quite sensitive to variation in $P(q)$ and consequently to differences in $SE[I_g(q)]$ and $SE[I_g(not\ q)]$. At this level, ODS predicts that in the low $P(q)$ condition there should be more q card selections than not q card selections and that in the medium and high $P(q)$ conditions there should be more not q card selections than q card selections. Qualitatively these predictions are supported by the CFIs in each condition except for the medium condition. On average, CFI was negative for the low and medium $P(q)$ conditions, indicating more q than not q card selections. In contrast, on average CFI was positive in the high $P(q)$ condition, indicating more not q than q card selections.

We further investigated these results by conducting a 3 × 2 mixed ANOVA with $P(q)$ as the between-subjects variable

Table 3

Mean Proportion of q Cards, Mean Proportion of not q Cards, and Consequent Falsification Index (CFI; Arcsine Corrected) for Each Level of P(q) in Experiment 3

	Mean prop. q		Mean prop. not q		CFI	
$P(q)$	M	SD	M	SD	M	SD
Low ($\frac{1}{8}$)	.69	.40	.63	.39	−0.12	2.17
Medium ($\frac{1}{2}$)	.75	.45	.56	.42	−0.53	2.42
High ($\frac{7}{8}$)	.30	.35	.89	.26	1.73	1.76

Note. Prop. = proportion; $P(q)$ = probability of q.

and card as the within-subjects variable and with proportions of cards selected (arcsine corrected) as the dependent variable. We used simple effects comparisons at each level of $P(q)$ to the test the above predictions. At the low and medium $P(q)$ levels, there were no significant differences between the proportion of q cards selected and the proportion of not q cards selected, $F(1, 33) < 1$, in both cases. However, at the high $P(q)$ level the proportion of not q cards selected was significantly higher than the proportion of q cards selected, $F(1, 33) = 7.91, MSE = 2.28, p < .01$.

Finally, at Level 2, ODS predicts the following interactions between these simple effects: There should be more q card selections and fewer not q card selections in the low $P(q)$ condition than in the medium and the high $P(q)$ conditions and more q card selections and fewer not q card selections in the medium $P(q)$ condition than in the high $P(q)$ condition (Prediction 2, Level 2). To test these pairwise interactions between conditions, we treated q and not q cards as levels of single within-subjects cards variable and performed three separate 2 × 2 mixed ANOVAs with proportion of cards selected (arcsine corrected) as the dependent variable. For the low versus high $P(q)$ comparison, there was a significant crossover interaction, $F(1, 22) = 5.30, MSE = 1.95, p < .05$, such that there was a higher proportion of q cards selected and a lower proportion of not q cards selected in the low $P(q)$ condition than in the high $P(q)$ condition. There was also a significant interaction for the medium versus high $P(q)$ comparison, $F(1, 22) = 6.87, MSE = 2.25, p < .025$, such that there was a higher proportion of q cards selected and a lower proportion of not q cards selected in the medium $P(q)$ condition than in the high $P(q)$ condition. However, the interaction between the medium and the low $P(q)$ conditions was not significant, $F(1, 22) < 1$. Apart from the final interaction, these results confirm Prediction 2, Level 2 of ODS. Moreover, they directly replicate the results of Experiment 1, even though in this experiment the simple effect for the low $P(q)$ condition was not replicated.

The results of Experiment 3 are consistent with the predictions of ODS and in the main replicate the findings of Experiment 1. In this experiment, participants selected from equal sized stacks of cards drawn from larger packs reflecting different probabilities of q cards. Consequently, there was no possibility of a small stack bias or that selecting cards at random could have generated these results. However, participants showed behavior very similar to that observed in Experiment 1. Therefore, it would appear that the results of Experiment 1 were not due to random sampling or to a small stack bias.

Experiment 3 replicated Experiment 1 in not confirming the Level 2 prediction of more not q than q card selections in the medium $P(q)$ condition. However, it did not replicate Experiment 1 insofar as it failed to confirm the Level 2 prediction of more q than not q cards selections in the low $P(q)$ condition. The Level 2 predictions depend on participants being quite sensitive to variation in $P(q)$. Consequently, one explanation for this pattern of results follows the possibility we mentioned in the discussion of Experiment 1. The failure to observe the predicted difference for the medium $P(q)$ condition may have occurred because

people require higher discriminability to override the default rarity assumption (see Figure 1).

The results for the low $P(q)$ condition may also be explained in similar terms. In the low condition, $P(q) = .167$, which is close to the region of uncertainty, even with the assumption that people are quite sensitive to $P(q)$ (see Figure 1B). Consequently, the failure to observe the Level 2 prediction for the low $P(q)$ condition in Experiment 3 may also be explained by a lack of sensitivity. However, the same lack of sensitivity was not observed for the low $P(q)$ condition in Experiment 1. We therefore explored these effects further in Experiment 4.

Experiment 4

Another possible explanation for the effects observed for the medium and low $P(q)$ conditions in Experiments 1 and 3 concerns sequential sampling. As we mentioned in introducing Experiment 1, ODS is about where to direct attention to pick up the most informative evidence. Accumulating evidence, as in sequential sampling in the RAST, may serve to redirect attention to evidence types other than those initially recommended by ODS. As we discuss in the General Discussion section, such redirection of attention may result from sequential sampling altering the parameters of the ODS model online. An immediate prediction is that participants' card selections should be as predicted by ODS for all conditions if the sequential sampling component of the task is removed. This suggests focusing on the initial card selected which can be influenced only by the initial parameter values and not by subsequent sampling. Experiments 1–3 focused on the percentage of cards selected from each stack as the dependent variable. This measure is also sensitive to participants refocusing attention on other types of evidence as sampling proceeds. Concentrating instead on the initial card selected provides a way of testing ODS in the absence of sequential sampling.

In Experiment 4, we therefore replicated and extended Experiment 3. We had participants perform two RASTs, one with the p, not p cards and one with the q, not q cards. For the q and not q cards task (QNQ task), we made the following predictions. If the discrepancy for the medium $P(q)$ condition is because of sensitivity, then there would be no differences in the frequency of initial selections of not q and q cards. Conversely, if the discrepancy is because of sequential sampling, the opposite result is predicted; that is, there would be significantly more initial selections of the not q card than the q card in the medium $P(q)$ condition. Furthermore, similar predictions hold for the low $P(q)$ condition. If the discrepancy between Experiments 1 and 3 for the low $P(q)$ condition is because of sensitivity, then there would be no differences in the frequency of initial card selections for the not q and the q cards. Conversely, if the discrepancy is because of sequential sampling, the opposite result is predicted; that is, there would be significantly more initial selections of the q card than the not q card in the low $P(q)$ condition. For the p and not p cards task (PNP task), ODS predicts that more participants would initially select the p card than the not p card across all conditions. This is

because the choice of initial card is mutually exclusive. Therefore, the card with the highest information gain, p, should win to the exclusion of the not p card.[11]

Method

Participants. There were 36 participants. All participants were recruited from the Department of Psychology, University of Warwick participant panel. Each participant was paid £2.00 for his or her participation. None of the participants had any prior knowledge of Wason's (1966) selection task.

Design. In this experiment, participants performed two RASTs. One RAST was identical to that in Experiment 3. The other RAST was the same as in Experiment 3 except that the p and not p card stacks were used as in Experiment 2. With three probability conditions and two possible orders in which to conduct the tasks, there were 18 possible combinations of task order and condition. With 36 participants, each possible combination was used twice.

Materials. The materials were the same as those in Experiment 3.

Procedure. The procedure was similar to that of Experiment 3, with the following changes. After having performed the first RAST, participants moved to a second table and performed the second RAST according to their condition and task order.

Results and Discussion

In this experiment, the dependent variable of interest was which card participants selected first. Table 4 shows these results for all three conditions. We analyzed the PNP and QNQ tasks separately. From Table 4, it is clear that the pattern of results was as predicted. In the QNQ task, although not q dominated q for the medium and high conditions, q dominated not q for the low condition. In the PNP task, p dominated not p in all conditions.

Within each task, each category is mutually exclusive, and hence we used the chi-square statistic to test the statistical significance of these results.[12] For the QNQ task, we partitioned the degrees of freedom by first comparing the medium and high conditions with each other and then comparing these two conditions collapsed with the low condition (for the logic of partitioning, see Siegel & Castellan, 1988). As is obvious from Table 4, the first partition indicated that the effects of the high and medium conditions were identical, $\chi^2_1(1, N = 36) = 0$. The second partition revealed the predicted crossover interaction between the medium and high conditions collapsed and the low condition, $\chi^2_2(1, N = 36) = 3.85, p < .025$ (one-tailed). Although there were more not q than q card selections in the medium and high conditions, there were more q than not q card selections in the low condition as predicted by ODS.

We tested the prediction of more not q than q card selections in the medium and high $P(q)$ conditions by using the chi-square goodness-of-fit statistic (Siegel & Castellan,

___

[11] This contrasts with the normal selection task where any or all of the cards may be chosen.

[12] For all 2 × 2 contingency tables reported here, Yate's (Camilli & Hopkins, 1978) correction for continuity was not used because not all the marginal totals were fixed.

1988). As predicted by ODS, there were significantly more *not q* than *q* initial card selections in the medium $P(q)$ condition, $\chi^2(1, N = 12) = 3.00, p < .05$ (one-tailed) and in the high $P(q)$ condition, $\chi^2(1, N = 12) = 3.00, p < .05$ (one-tailed) than would be expected by chance. However, casual inspection shows that there were no more *q* than *not q* card selections in the low $P(q)$ condition. Note that the selections of *not q* and *q* cards were the same in the medium and the high $P(q)$ conditions. However, ODS predicts that there should be more *not q* and less *q* card selections in the high $P(q)$ condition than in the medium $P(q)$ condition. We offer no explanation for this finding apart from noting that it was probably due to chance variation and that perhaps a larger sample size may confirm this prediction for the initial card selections in the same way that Experiments 1 and 3 confirmed it for all card selections.

For the PNP task, we tested to what extent selections differed from chance by using the chi-square goodness-of-fit statistic (Siegel & Castellan, 1988). As predicted, for each condition the distribution of *p* and *not p* card selections differed significantly from chance: low, $\chi^2(1, N = 12) = 5.33, p < .025$ (one-tailed); medium, $\chi^2(1, N = 12) = 8.33, p < .005$ (one-tailed); high, $\chi^2(1, N = 12) = 3.00, p < .05$ (one-tailed).

For completeness, we also show the results of Experiment 4 presented as in Experiments 1–3. We show the proportion of each card selected by order and by condition in Table 5. Because of a strong order effect, the results of Experiment 3 were replicated—albeit not significantly—only when the QNQ task was performed first. Order affected the consequent card selections (*q* and *not q*) such that in the PNP–QNQ order, there were more *not q* than *q* cards selected for all conditions, but in the QNQ–PNP order, there were more *not q* than *q* cards selected for the high $P(p)$ condition only, as in Experiments 1 and 3. In a 2 × 3 between-subjects ANOVA with order and condition as variables and proportion of *q* cards selected as the dependent variable, there was a significant main effect of order, $F(1, 30) = 6.24, MSE = 0.16, p < .025$, such that there was a higher proportion of *q* card selections in the QNQ–PNP order than in the PNP–QNQ order.

Could these order effects explain the significant results found for the initial card selections? The main order effect was for high levels of *not q* card selections for all conditions and for fewer *q* card selections in the PNP–QNQ order. If anything, this would predict more *not q* than *q* card selections in all conditions in Experiment 4. However, we still observed the predicted interactions between conditions

Table 4
*Number of Cards Initially Chosen From the Card
Stacks in Experiment 4*

| $P(p)$ or $P(q)$ | Card stack | | | |
	p	*not p*	*q*	*not q*
Low ($\frac{1}{6}$)	10	2	7	5
Medium ($\frac{1}{2}$)	11	1	3	9
High ($\frac{5}{6}$)	9	3	3	9

Note. $P(p)$ = probability of *p*; $P(q)$ = probability of *q*.

Table 5
*Proportion of Cards Selected in Each Order and in Each
Condition in Experiment 4*

| | Condition | | | | | |
| | Low | | Medium | | High | |
Card and order	M	SD	M	SD	M	SD
p						
PNP–QNQ	.92	.20	1.00	.00	.77	.38
QNQ–PNP	.87	.22	1.00	.00	1.00	.00
not p						
PNP–QNQ	.35	.44	.10	.20	.35	.51
QNQ–PNP	.33	.52	.35	.42	.08	.12
q						
PNP–QNQ	.47	.44	.28	.40	.52	.46
QNQ–PNP	.70	.38	.93	.16	.62	.45
not q						
PNP–QNQ	.83	.32	.83	.41	.85	.37
QNQ–PNP	.70	.34	.75	.42	.83	.26

Note. PNP = the *p* and *not p* cards task; QNQ = the *q* and *not q* cards task.

for initial card selections. Consequently, the order effect in Experiment 4 could not explain the results for initial card selections because it would predict the opposite result to that observed. We explore possible reasons for these order effects in the General Discussion.

The results of Experiment 4 are largely consistent with ODS. Participants' attention was initially focused on the cards that ODS predicts. However, in the low condition, it would appear that $P(q)$ was close to the area of uncertainty, which is why no significant differences between *q* and *not q* card selections were observed in this experiment and in Experiment 3. As we discuss in the General Discussion, this difference with Experiment 1 for the low $P(q)$ condition may be understood in terms of the effects of sequential sampling and the different stack sizes used in that experiment. Moreover, in the medium $P(q)$ condition, it would appear that participants subsequently attended to classes of evidence other than those initially recommended by ODS alone. (Although as we argue in the General Discussion, if the model's parameters are being updated online, then the ODS model may predict these results.) Because of the binary choice at each trial in the RAST, any effect of accumulating evidence must be to focus attention away from the cards ODS initially recommends. In spite of this, the trends predicted by ODS were significantly observed in the RAST in Experiments 1–3. Thus, it would appear that if additional mechanisms are needed to explain refocusing (which they may not be, see the General Discussion), then they overlay the basic mechanisms of ODS. It is important that when the effects of refocusing were eliminated by looking only at the first card chosen, as we did in Experiment 4, the predictions of ODS were again confirmed.

General Discussion

The purpose of these experiments was to test the predictions of Oaksford and Chater's (1994) ODS model. In all these experiments, the parameters of the model were varied

at three levels—low, medium, and high—by varying the sizes of the card stacks used in the RAST. Predictions were derived at two different levels of sensitivity to variation in the model's parameters, $P(q)$ and $P(p)$. At Level 1, where participants were assumed to be relatively insensitive to such variation, four predictions were derived at increasing levels of detail. The Level 1 predictions were all between conditions. At Level 2, where participants were assumed to be quite sensitive to varying the model's parameters, predictions were also made within conditions concerning the relative frequency of different cards that should be selected. In summary, Experiment 1 confirmed all the Level 1 predictions of the ODS model. The experiment confirmed all the Level 2 predictions except that there should be more *not* q than q card selections for the medium $P(q)$ condition and the related prediction that there should be an interaction for the proportions of q and *not* q cards selected for the medium versus low $P(q)$ conditions.

The results of Experiment 1 could be explained by participants selecting at random or by a small stack bias; in other words, participants may simply have tended to choose a higher proportion of cards from small stacks. Experiment 2 controlled for these possible confounds by using a RAST with the p and *not* p cards. Random selection or small stack bias would predict trends similar to those observed in Experiment 1—that participants should select p cards in preference to *not* p cards for a small p stack and *not* p cards in preference to p cards for a large p stack with intermediate choices for a medium p stack. ODS, by contrast, predicts that the p card should always be preferred. Furthermore, it predicts that the *not* p selections will be greatest for a large p stack, but because ODS predicts that the p card should be so much preferred in all conditions, the termination conditions of the RAST mean that p selections will always be at ceiling. Experiment 2 provided strong evidence against random selection and small stack bias. As ODS predicts, p card selections were significantly higher than *not* p card selections for all levels of $P(p)$. Moreover, the predicted trend for *not* p card selections to increase as $P(p)$ increases was very close to significance.

However, it could be argued that although random sampling and small stack bias are not factors for the p cards, they might nonetheless affect performance in the original RAST with q and *not* q cards. Experiment 3 removed both possibilities by using stacks of equal size in all conditions and manipulating probabilities by changing the size of the packs from which the stacks were drawn. ODS makes the same predictions in Experiment 3 as in Experiment 1, which Experiment 3 largely replicated at Level 1. It also replicated Experiment 1 at Level 2 insofar that in the medium $P(q)$ condition, no difference was observed in the frequency of q and *not* q card selections. However, Experiment 3 did not replicate Experiment 1 insofar as the same result was found for the low $P(q)$ condition. Neither result is consistent with the Level 2 predictions of the ODS model.

Although these Level 2 results could be attributed to a simple lack of sensitivity, we investigated them further in Experiment 4. We argued that the discrepancies for the low and medium conditions could be due either to a lack of

sensitivity or to sequential sampling leading to changes of attentional focus to evidence other than that initially recommended by ODS. In Experiment 4, we therefore looked at the first card selected which, not being susceptible to any subsequent changes of attentional focus, provides a more critical test of our model's predictions. For the medium $P(q)$ condition, Experiment 4 confirmed the predictions of ODS—the discrepancy for this condition observed in Experiments 1 and 3 disappeared for the first card selected. This result suggests that the findings for this condition in Experiments 1 and 3 were due to sequential sampling. However, for the low $P(q)$ condition, the results of Experiment 4 mirrored those of Experiment 3. This result suggests that the low condition, where $P(q) = .167$, is in the region of uncertainty (see Figure 1B). As we noted in introducing the predictions of the ODS model, even if people are quite sensitive to manipulating its parameters, $P(q) = .167$ will be close to this region. The significant Level 2 result for the low $P(q)$ condition in Experiment 1 may have been due to chance, but we offer an alternative explanation below in terms of sequential sampling.

Experiments 1–4 confirmed that there are significant effects of varying the parameters of the ODS model in the RAST. The results of these experiments are broadly in line with the predictions of the ODS model down to quite a fine level of detail. It is important to note that even if the precise predictions of the ODS model were not confirmed in these experiments, the significant effects we observed are not consistent with most other major theoretical accounts of selection task performance. We now show this in a review of how other theoretical approaches may attempt to explain these data.

First, we deal briefly with theories that seem in principle incapable of handling these results. Mental logic (e.g., Braine, 1978; Henle, 1962; Rips, 1983, 1994)—the proposal that we reason in these tasks by using syntactic rules or *proof theory*—provides no mechanisms for handling probabilistic effects. However, as Oaksford and Chater (1994) observed, Rips (1990) has argued that the selection task is not a deductive reasoning task but is rather a "loose" probabilistic task. Consequently, although mental logic cannot explain these effects, they do not necessarily question the utility of the mental logic approaches to "tight" logical tasks where it may provide a more appropriate approach (although see Oaksford & Chater, 1995b, for discussion of whether such tasks exist outside of mathematics).

There have been some recent proposals for how to extend mental-models theory to handle probabilistic effects (Johnson-Laird, 1994). However, it is not clear how this extension could be applied in particular cases. Moreover, the core of mental models as an approach to inference has been the claim that reasoning is the search for counterexamples. But when it comes to probabilistic relations, counterexamples will always be possible. Consequently, the very basis on which the mental models theory of inference operates must be revised to account for probabilistic effects like those in Experiments 1–4. Further, recent work on focusing and mental models may be thought to explain why participants refocus on cards other than those initially recommended by ODS (Legrenzi, Girotto, & Johnson-Laird, 1993). The

suggestion is that people focus only on the contents of their mental models, but this just restates the problem. We want to know what evidence people attend to; no doubt this is the evidence that people then represent, but what makes them attend to it in the first place? Chater and Oaksford (1993; see also, Oaksford & Chater, 1995b) argued that there is nothing intrinsic to mental models theory that can provide an answer to this question.

Sperber et al. (1995) have suggested that relevance theory (Sperber & Wilson, 1986) can account for attentional changes in the selection task. Consistent with relevance theoretic accounts, Oaksford and Chater (1995a) have proposed that ODS provides a formal relevance measure that serves to direct attention to informationally salient evidence. Consequently, these accounts are not in competition. However, although agreeing on the importance of relevance processing in reasoning (see also Evans, 1994), Sperber et al.'s relevance account could not explain the results of Experiments 1–4. Their account is framed in terms of cognitive effects and effort that do not appear to have any immediate application to the purely probabilistic manipulations used in these experiments. Indeed, Oaksford and Chater (1995a) argued that the success of Sperber et al.'s experiments demonstrating relevance theoretic effects may have been due to their implicitly manipulating $P(p)$ and $P(q)$. We now consider approaches that contain an explicit probabilistic component.

Kirby's (1994) model suggests that people attend to the cards in the selection task depending on the probability of finding a counterexample. Consequently, when $P(p)$ is low, participants should be less inclined to check the *not q* card because it is unlikely to be a *p, not q* instance. However, when $P(p)$ is high, the *not q* card is highly likely to yield such a counterexample. Kirby's emphasis on searching for counterexamples means that he was unable to explain why we also observed significant changes in selections for the *q* card in these experiments (a result that Kirby also found). These changes are predicted by Oaksford and Chater's (1994) Bayesian model because it allows that both confirming and disconfirming evidence is relevant. In, for example, Experiment 3, the difference in *q* card selections between the high and low conditions was significant, $F(1, 33) = 5.65$, $MSE = 16.30, p < .025$, despite the tendency to also select *not q* cards in the low condition. Kirby's model cannot explain these effects.

Recently, Over and Evans (1994, see also Evans & Over, 1996) have proposed an account based on what they call *epistemic utility*. They have not provided a formal model of this account so it is difficult to apply in particular cases. The principal contrast with Oaksford and Chater's (1994) account is that different classes of evidence may have different epistemic values in different contexts. Consequently, rather than laying all the emphasis on counterexamples as in Kirby's (1994) model or allowing that confirming and disconfirming evidence are weighted depending solely on $P(p)$ and $P(q)$ as in Oaksford and Chater's (1994) model, Over and Evans argued that confirmation and disconfirmation may also be differentially weighted depending on a person's goals. Although this is an eminently reasonable suggestion

as Oaksford and Chater (1996) observed, without a formally specified model we are unable to assess whether Over and Evan's proposal would make different predictions than ODS did in Experiments 1–4. Nevertheless, there are aspects of our data that may point to an epistemic utility account.

The finding that during sequential sampling participants' attention may have been drawn to cards other than those initially recommended by ODS is open to a variety of interpretations. First, consistent with Oaksford and Chater's (1994) current ODS model, it could be that participants were updating $P(p)$ and $P(q)$ online as a result of sequential sampling. Although participants were told the distribution of the *q* (or *p*) cards at the outset and these values determined their initial selections as Experiment 4 revealed, they may nonetheless have treated the stack sizes as only tentative estimates of $P(p)$ and $P(q)$ to be revised when sampling the real data. We used continuous updating of $P(p)$ and $P(q)$ in a simulation, and it would indeed appear that this can lead to trial-by-trial alternations in the stacks that ODS recommends as the most informative.

Second, consistent with Over and Evans's (1994) article, it could be that sequential sampling was altering participants' goals. According to ODS, in both the medium and high $P(q)$ conditions, participants' attention was initially directed toward the *not q* card stack. All the cards in this stack were *not p, not q* cards; that is, although examining these cards would confirm the rule—that is, $P(M_D)$ rises— they provided participants with no positive instances of the rule. This is like arguing that although you have never seen a raven, you are confident that all ravens are black because all the nonblack things you have examined were also not ravens. As discussed earlier, the ravens paradox is eliminated by Oaksford and Chater's (1994) default rarity assumption (Horwich, 1982; Howson & Urbach, 1989). In the low $P(q)$ condition, participants' attention was focused on the *q* cards, which were all *p, q* instances; that is, participants were looking for blue triangles (black ravens). However, when rarity was not in force (i.e., in the medium and high $P(q)$ conditions), participants may have wanted to see some blue triangles as well as red circles to confirm that the rule was not just vacuously true. Consequently, although participants were initially guided to the *not q* stack in the medium and high $P(q)$ conditions by ODS, they may at some time have decided to search for positive instances of the rule.

However, participants were more likely to look for positive instances in the medium condition than in the high condition. $SEI_g(not\ q)$ was much higher in the high $P(q)$ condition than in the medium $P(q)$ condition. Consequently, it would take more data to overcome the attentional focus on the *not q* card in the high $P(q)$ condition than in the medium $P(q)$ condition. However, in the RAST the number of data points are limited to a fixed number. (In Experiments 3 and 4, it was 10.) Therefore, the point where attention is refocused was more likely to have been met in the medium $P(q)$ condition than in the high $P(q)$ condition.

Conversely, in the low $P(q)$ condition after initially focusing on the *q* cards, participants' attention may have been diverted toward the *not q* cards. From our Bayesian perspective, both *p, q* and *p, not q* cases may provide

important classes of evidence. After examining a certain number of p, q cases, attention may have shifted to looking for p, not q cases. This could explain the differences between Experiments 1 and 3 in the low condition. In Experiment 1, in the low $P(q)$ condition there were 5 q cards, whereas in Experiment 3 there were 10. It was therefore more likely that participants would exhaust this stack while their attention was still focused on p, q instances in Experiment 1 than in Experiment 3. Because these instances were then exhausted, participants declared the rule true and terminated the experiment. Consequently, a far lower proportion of not q cards should have been selected in the low $P(q)$ condition in Experiment 1 than in Experiment 3, which would be consistent with the observed differences between these experiments.[13]

This explanation for attentional shifts during sequential sampling can also explain the order effect observed in Experiment 4. In the PNP-QNQ order, participants initially chose all the p cards, which all had qs on the other side. They then moved on to the QNQ task. However, by then they had already observed up to 10 positive p, q instances. Consequently, they focused attention on the p, not q cases. However, in the QNQ-PNP order, participants did not have the same opportunity to examine positive instances of the rule.

Current epistemic utility theory does not address how these changes of attention come about. However, a simple proposal would be to allow different utilities of evidence in Oaksford and Chater's (1994) ODS model (as Evans & Over, 1996, suggest) in a manner similar to their maximum expected utility model of the deontic selection task (Oaksford & Chater, 1994, pp. 621–625). Modeling changes of attention due to sequential sampling would involve making the utilities for each evidence type some decreasing function of the number of instances of that evidence type observed. If all evidence were equally weighted initially, such a model would make the same predictions as ODS for initial card selections. However, after accumulating data, other evidence types would come to have higher utility as the utility of the evidence type initially recommended by ODS falls off.

It is important to note that these proposals are consistent with the ODS framework. If people were doing the mental equivalent of updating their parameter values online, then they were effectively changing the models (i.e., M_I and M_D) that are compared in ODS. As with the changes in these models we intended to bring about in these experiments by explicitly manipulating $P(q)$ and $P(p)$, the ODS model still predicts that participants select the optimally informative experiments to perform. However, as the models change, which experiments are optimal may also change. A similar argument applies to the suggestion that utilities of different evidence types should be introduced into these calculations as seems to be suggested by epistemic utility theory. We suggested that such an approach could be incorporated by introducing a utility function with respect to evidence types that is some decreasing function of the number of instances observed. Weighting these utilities by expected information gain provides a quantity that it would be worth maximizing in the same way as any other expected utility measure. Consequently, any of the revisions to the ODS model that we

have suggested in response to Experiments 1–4 are still very much in the spirit of ODS.

Which of these possible accounts of the observed deviations from ODS is the right one must await further research. The RAST provides an excellent paradigm for investigating these issues. This investigation will involve detailed analyses of the sequences of selections that people make that we did not investigate in these experiments. It will further involve simulating the various models, like epistemic utility and the revised ODS model, to see if they can be discriminated by virtue of their predictions for the structure of sequential samples. It is important, however, to note that although we did observe some deviations from ODS, the data reported here are most consistent with that model.

---

[13] This would also be true because the number of not q cards in Experiment 1 was 25, whereas it was 10 in Experiment 3.

References

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

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

Anderson, J. R. (1994). *Rules of the mind.* Hillsdale, NJ: Erlbaum.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85*, 1–21.

Camilli, G., & Hopkins, K. D. (1978). Applicability of chi-square to 2 × 2 contingency tables with small expected frequencies. *Psychological Bulletin, 85*, 163–167.

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

Chater, N., & Oaksford, M. (1993). Logicism, mental models and everyday reasoning: Reply to Garnham. *Mind & Language, 8*, 72–89.

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

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioural & 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.

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

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

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

Evans, J. St. B. T. (1994). Relevance and reasoning. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning* (pp. 147–171). Hillsdale, NJ: Erlbaum.

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. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review, 103*, 356–363.

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

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.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43,* 127–171.

Girotto, V. (1988). Pragmatic knowledge and deductive reasoning in children. *Giornale Italiano di Psicologia, 15,* 287–314.

Girotto, V., Blaye, A., & Farioli, F. (1989). A reason to reason: Pragmatic basis of children's search for counterexamples. *Cahiers de Psychologie Cognitive, 9,* 297–321.

Girotto, V., Light, P., & Colbourn, C. J. (1988). Pragmatic schemas and conditional reasoning in children. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 40*(A), 469–482.

Goodman, N. (1983). *Fact, fiction and forecast.* Cambridge, MA: Harvard University Press. (Original work published 1954)

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

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69,* 366–378.

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

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

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition, 50,* 189–209.

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

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

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

Keynes, J. M. (1921). *A treatise on probability.* London: Macmillan.

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

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

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 551–578.

Light, P., Blaye, A., Gilly, M., & Girotto, V. (1989). Pragmatic schemas and logical reasoning in 6- to 8-year old children. *Cognitive Development, 4,* 49–64.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1979). *Introduction to bivariate and multivariate analysis.* Glenview, IL: Scott, Foresman.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of mathematical statistics, 27,* 986–1005.

Mackay, D. J. C. (1992). *Bayesian methods for adaptive models.* Unpublished doctoral dissertation, California Institute of Technology, Pasadena, CA.

Manktelow, K. I., & Over, D. E. (1987). Reasoning and rationality. *Mind & Language, 2,* 199–219.

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.

McNicol, D. (1972). *A primer of signal detection theory.* London: Allen & Unwin.

Mises, R. von (1939). *Probability, statistics and truth.* London: Allen & Unwin.

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

Oaksford, M., & Chater, N. (1992). Bounded rationality in taking risks and drawing inferences. *Theory & Psychology, 2,* 225–230.

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

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

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

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

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

Pollard, P. (1985). Nonindependence of selections on the Wason selection task. *Bulletin of the Psychonomic Society, 23,* 317–320.

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. (Original work published 1935)

Pylyshyn, Z. W. (1984). *Computation and cognition.* Cambridge, MA: MIT Press.

Ramsey, F. P. (1931). *The foundations of mathematics and other logical essays.* London: Routledge and Kegan Paul.

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. (1988). *Non-parametric statistics for the behavioral sciences.* New York: McGraw Hill.

Snodgrass, J. G. (1977). *The numbers game: Statistics for psychology.* New York: Oxford University Press.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition, 57,* 31–95.

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

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

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

Thagard, P. (1988). *Computational philosophy of science.* Cambridge, MA: MIT Press.

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Harmonsworth, Middlesex, England: Penguin Books.

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

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

Wason, P. C., & Green, D. W. (1984). Reasoning and mental representation. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 36*(A), 597–610.

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

Wiener, N. (1948). *Cybernetics.* New York: Wiley.

Wolfram, S. (1991). Mathematica [Computer program]. Reeding, MA: Addison-Wesley.

American Psychological Association
Subscription Claims Information

Today's Date: _____

We provide this form to assist members, institutions, and nonmember individuals with any subscription problems. With the appropriate information we can begin a resolution. If you use the services of an agent, please do NOT duplicate claims through them and directly to us. **PLEASE PRINT CLEARLY AND IN INK IF POSSIBLE.**

PRINT FULL NAME OR KEY NAME OF INSTITUTION

ADDRESS

CITY   STATE/COUNTRY   ZIP

YOUR NAME AND PHONE NUMBER

TITLE

MEMBER OR CUSTOMER NUMBER (MAY BE FOUND ON ANY PAST ISSUE LABEL)

DATE YOUR ORDER WAS MAILED (OR PHONED)

____PREPAID ____CHECK ____CHARGE
              CHECK/CARD CLEARED DATE:_____

(If possible, send a copy, front and back, of your cancelled check to help us in our research of your claim.)

ISSUES: ____MISSING ____DAMAGED

VOLUME OR YEAR   NUMBER OR MONTH

Thank you. Once a claim is received and resolved, delivery of replacement issues routinely takes 4–6 weeks.

━━━━━━━━ (TO BE FILLED OUT BY APA STAFF) ━━━━━━━━

DATE RECEIVED: _____   DATE OF ACTION: _____
ACTION TAKEN: _____   INV. NO. & DATE: _____
STAFF NAME: _____   LABEL NO. & DATE: _____

Send this form to APA Subscription Claims, 750 First Street, NE, Washington, DC 20002-4242

PLEASE DO NOT REMOVE. A PHOTOCOPY MAY BE USED.