Psychonomic Bulletin & Review
2004, 11 (3), 521-527

Logical reasoning and probabilities: A comprehensive test of Oaksford and Chater (2001)

KLAUS OBERAUER, ANDREA WEIDENFELD, and ROBIN HÖRNIG
University of Potsdam, Potsdam, Germany

We report two experiments testing a central prediction of the probabilistic account of reasoning provided by Oaksford and Chater (2001): Acceptance of standard conditional inferences, card choices in the Wason selection task, and quantifiers chosen for conclusions from syllogisms should vary as a function of the frequency of the concepts involved. Frequency was manipulated by a probability-learning phase preceding the reasoning tasks to simulate natural sampling. The effects predicted by Oaksford and Chater (2001) were not obtained with any of the three paradigms.

For several decades, psychological research on deductive reasoning and research on probabilistic reasoning have largely been separate. Recently, there has been increasing awareness that probabilities might matter also in tasks framed as deductive inferences (see, e.g., Evans & Over, 1996a; Kirby, 1994; Nickerson, 1996). One of the most elaborated theories highlighting the role of probabilistic reasoning in deductive tasks has been advanced by Oaksford and Chater (1994, 2001). In this article, we report two experiments testing central predictions of this theory for three paradigms: the Wason selection task, conditional inferences, and syllogisms.

The Wason Selection Task

Wason's (1966) task requires selection of information necessary to test a conditional of the form "if *p* then *q*." Four cards are given, each representing a combination of *p* or *not-p* with *q* or *not-q*. The visible sides of the cards correspond to the four logically possible cases: *p*, *not-p*, *q*, and *not-q*, respectively. In order to test whether the conditional, interpreted as a material implication, is true for the four cards, one has to turn over the cards displaying *p* and *not-q*. Few participants, however, choose this combination for task versions with nonthematic materials (see Evans, Newstead, & Byrne, 1993).

Oaksford and Chater (1994) developed a theory of people's choices in the selection task based on the expected information gain (EIG) associated with turning each card. Their optimal data selection (ODS) theory interprets the selection task as an inductive, not a deductive, task: Participants understand that they should test a universal law, not a statement about the four cards shown; their task, therefore, consists of selecting those cards that promise to be most informative with regard to evaluating the universal hypothesis.

The result of Oaksford and Chater's (1994) analysis is that the relative EIG of the cards depends on the assumptions about the proportions of *p* to *not-p* cases and of *q* to *not-q* cases in the population. The default assumption of the cognitive system, they argue, is *rarity*: Both *p* and *q* are less frequent than their negations. Under the assumption of rarity, the *p* card would be most informative, and the *q* card would come next. Thus, the modal pattern of selections, *p* and *q*, would be rational.

The most important prediction of ODS is that the probability of selecting individual cards should vary as a function of the relative frequencies of *p* and *q*. The main predicted effects of ODS in its revised version (Oaksford & Chater, 2001) are that the selection probability of *p* decreases and that of *not-p* increases as the frequency of *p* increases. Likewise, the selection probability of *q* decreases and of *not-q* increases as the frequency of *q* increases (see Table 1). In other words, ODS predicts that people tend to select cards from the rare categories.

Empirical tests of this prediction are sparse and inconclusive. Oaksford, Chater, Grainger, and Larkin (1997) gained support for predictions from ODS with the reduced-array selection task (RAST). The RAST is a simplified version of Wason's original task, and it is not clear whether the results can be generalized to the original task (cf. Oberauer, Wilhelm, & Rosas Diaz, 1999). Oaksford, Chater, and Grainger (1999) varied the probabilities of *p* and of *q* in the Wason selection task. Over four experiments, they obtained partial support for the effects predicted by ODS. In each experiment, however, some of the predicted effects were not obtained. Disturbingly, the successes and failures of predictions varied from one experiment to the next without a clear pattern. Oberauer et al. (1999) also varied the probabilities of *p* and *q*. They found weak support for ODS in the first experiment and no ef-

This research was supported by DFG Grant FOR 375 (B2). We thank Petra Grüttner and Artur Schneider for help with collecting the data. Correspondence concerning this article should be addressed to K. Oberauer, Department of Psychology, University of Potsdam, P.O. Box 60 15 53, 14415 Potsdam, Germany (e-mail: ko@rz.uni-potsdam.de).

Table 1
Predicted Trends for Selections in the Wason Selection Task and Acceptances in Conditional Inferences, Together With Results of the Experiment

Decision	Predicted higher for	Contrast	Contrast Value, Experiment 1 (*t* Value)	Contrast Value, Experiment 2 (*t* Value)
		Wason Selection Task		
Selection of				
Object *p*	small *P*(*p*)	(LL + LH)–(HL + HH)	−0.13 (−1.42)	−0.31 (−3.05)
Object *not-p*	large *P*(*p*)	(HL + HH)–(LL + LH)	0.01 (0.22)	−0.32 (−4.06)
Object *q*	small *P*(*q*)	(LL + HL)–(LH + HH)	0.03 (0.31)	−0.10 (−0.91)
Object *not-q*	large *P*(*q*)	(LH + HH)–(LL + HL)	0.06 (0.85)	−0.19 (−2.16)
		Conditional Inferences		
Acceptance of				
DA	small *P*(*q*)	(LL + HL)–(LH + HH)	−0.13 (−1.45)	0.03 (0.31)
AC	large *P*(*p*)	(HL + HH)–(LL + LH)	0.15 (1.80)	0.08 (0.77)
MT	small *P*(*p*)	(LL + LH)–(HL + HH)	−0.12 (−1.27)	−0.17 (−1.87)

Note—DA, denial of the antecedent; AC, acceptance of the consequent; MT, modus tollens. LL, LH, HL, and HH designate the experimental conditions formed by crossing the probability of *p* (first letter: L = low, H = high) with the probability of *q* (second letter).

fect of the probability manipulation in two subsequent experiments. Handley, Feeney, and Harper (2002) also found no effect of a manipulation of the probability of *q*.

The experiments by Oberauer et al. (1999) could be criticized because they introduced frequency information about *p* and *q* verbally as part of the instructions. Although it was checked that participants understood this information, they might not have processed it in a way that could override people's default rarity assumption. Gigerenzer and Hoffrage (1995) argued that evolution has shaped our minds to process probabilistic information best when it is presented in *frequency formats* resulting from *natural sampling*. Natural sampling means that the mind keeps counters of events from various categories experienced one by one and estimates probabilities by comparing the counters.

Following this line of argument, Oaksford and Wakefield (2003) conducted an experiment on a variant of the selection task in which the participants saw 40 cards, one by one, and decided for each card whether they would have to turn it to check whether the conditional was true. The probability of *p* (vs. *not-p*) and of *q* (vs. *not-q*) was manipulated through the proportions of cards. In agreement with the predictions from ODS, the participants decided to turn a larger proportion of the rare cards. Unfortunately, this experiment confounded the frequency manipulation with the number of opportunities to turn cards from each category: The participants received only 2 cards from the rare categories but 18 from the frequent ones. If the participants had a tendency to avoid selecting redundant cards (i.e., cards from the frequent categories after their first two or three exemplars), this would reduce the proportion of cards selected from the frequent categories and thereby generate the predicted results.[1]

Here, we present two experiments in which a natural sampling procedure was used to introduce relative frequencies of categories. We unconfounded frequency and selection opportunity by having participants first learn the frequency distributions and then conduct Wason selection tasks separately.

Conditional Inferences

Oaksford, Chater, and Larkin (2000) formulated a probabilistic account for inferences from conditional premises. There are four basic inference forms, all of which take a conditional of the form "if *p* then *q*" as the major premise and one of four elementary propositions as the minor premise (see Table 2).

According to Oaksford et al. (2000), the conditional premise is understood as expressing a dependency between *p* and *q*, such that most, but not necessarily all, cases of *p* are cases of *q*. A person's tendency to endorse one of the inferences is a function of the conditional probability of the conclusion, given the minor premise. This can be calculated from three parameters: the probability of an exception to the conditional (i.e., of a *p*, *not-q* case) and the probabilities of *p* and *q*. The tendency to accept modus ponens (MP) depends only on the exception parameter. More important in the present context, the tendency to endorse acceptance of the consequent (AC) is predicted to increase with the probability of *p*; acceptance of denial of the an-

Table 2
Four Conditional Inference Forms With "If *p* Then *q*" as Major Premise

Inference Form	Modus Ponens	Denial of the Antecedent	Acceptance of the Consequent	Modus Tollens
Minor premise	*p*	*not-p*	*q*	*not-q*
Conclusion	*q*	*not-q*	*p*	*not-p*
Logical validity	yes	(only for biconditional)	(only for biconditional)	yes

tecedent (DA) is predicted to increase with increasing probability of *not-q*, and modus tollens (MT) should be endorsed more when the probability of *not-p* increases. In other words, conclusions should be drawn more readily if their prior probability is higher. Oaksford et al. (2000) reported three experiments yielding results in accordance with their predictions.

Syllogisms

Aristotelian syllogisms consist of two quantified statements as premises, from which a third quantified statement should be concluded (if possible) relating the two *end terms* of the premises (i.e., the terms not connected directly by one of the premises). For example, *some A are B*, *all B are C* licenses the conclusion *some A are C*. A probabilistic theory for syllogisms, the probability heuristics model (PHM), has been formulated by Chater and Oaksford (1999). Assuming again that the cognitive system takes rarity of the categories involved as default, the quantifiers can be ordered in terms of how informative they are, with *all* being the most informative, followed by *some*, *no*, and *some . . . not*.

The theory then states two heuristics by which the quantifier of a conclusion is selected. The *min-heuristic* says that the less informative quantifier from one of the premises is chosen for the conclusion. The *p-entailment heuristic* states that the next most preferred quantifier is the one entailed probabilistically by the quantifier selected through the *min-heuristic* (i.e., *all* entails *some*, *no* entails *some . . . not*, and *some* and *some . . . not* entail each other). Moreover, participants avoid drawing very uninformative conclusions—that is, conclusions with *some . . . not*.

Our test of PHM rests again on varying the degree to which the rarity assumption holds. *All A are B* is more informative than *no A is B* only as long as *B* is less frequent than *not-B*. The same holds for *some A are B* relative to *some A are not-B*. This is evident when we note that *no A is B* is equivalent to *all A are not-B* (and analogously for *some* and *some . . . not*). In general, conclusions with negative quantifiers (*no* and *some . . . not*) become more informative, and those with positive quantifiers (*all* and *some*) less informative, as the frequency of the terms they interrelate increases.

The syllogisms used in our study are listed in Table 6 below. The first six syllogisms combine premises that switch their informativeness ranks when the terms involved go from rare to frequent (the remaining two were fillers). PHM therefore should predict by the *min-heuristic* that *no* and *some . . . not* are preferred as quantifiers in the conclusions when *A*, *B*, and *C* are rare, but *all* and *some* are preferred when they are frequent and their negations are rare. The effect might not be that clear-cut if the cognitive system sticks to the rarity assumption as a default, but a strong frequency manipulation should have at least some effect in the expected direction.

METHOD

Participants

The participants were 68 (Experiment 1) and 72 (Experiment 2) high school and university students with a minimum age of 17 years. They received 5 Euro or partial course credit for one session, which lasted about 45 min.

Materials and Procedure

Both experiments were identical, except for one sentence in the instructions (explained below). Each experiment had five phases. Phase 1 consisted of a probability-learning task. In Phases 2–4, the participants worked on the reasoning tasks, beginning with four Wason selection tasks, followed by 16 conditional inferences and 16 syllogisms. Finally, the participants were asked to rate the probabilities learned at the beginning as a manipulation check. The whole procedure was computerized, and the computer displayed the instructions for each phase immediately before the phase started. Within each task phase, the participants could switch back to the instruction display of that phase.

The materials of all five phases consisted of four kinds of objects presented as simple sketches on the computer screen. Each object had three features with binary values. One of the features could be seen from both sides of the object, whereas each of the other two features was visible from one side only (see Table 3).

In the first phase, the participants went through a series of 240 guessing trials. Each trial started with a question specifying an object and a feature—for example, "Coin: symbol"—below which the two alternatives were displayed. The participants selected their guess by pressing the left or right arrow key, after which a picture of the object appeared, confirming or disconfirming the guess. For a correct guess, the participant gained one point, and the current count of points was always displayed below the picture of the object.

The objects and the feature dimensions to be guessed, as well as the correct feature values, were determined at random for each trial. The critical experimental variation was that for each feature dimension of the four objects, one value had a probability of appearing of .9, and the other of .1. Which values would be the frequent ones was determined at random for each participant at the beginning of the experiment.

In the second phase, the participants were presented with four trials of the Wason selection task, each involving a different object. The instruction made it clear that the conditional to be tested referred to "the objects you just familiarized yourself with." The conditional used the two features visible from only one side as the antecedent and the consequent. In this way, we could display pictures of four objects, so that two of them showed the values of the antecedent feature (one *p*, one *not-p*) and the other two the values of the conse-

Table 3
Objects and Their Features as Used in the Experiments

Object	Feature Visible on Both Sides	Feature on Side 1	Feature on Side 2
Playing card	form (round or sharp edges)	number (1 or 2)	letter (A or B)
Coin	color of border (yellow or violet)	value (10 or 50 cents)	symbol (half-moon or star)
Letter	form (square or elongated)	city in address (Hofstadt or Lohausen)	symbol on seal (% or $)
Cushion	form (square or round)	color (red or blue)	texture (dotted or hatched)

quent feature (one *q*, one *not-q*). Over the four trials, we crossed the frequency of *p* and the frequency of *q* by selecting the appropriate feature value for the antecedent and the consequent of the rule. The assignment of experimental conditions to the objects, the assignment of features to the antecedent and the consequent, and the order of the four tasks were determined at random for each participant.

Within each trial, the four objects were displayed simultaneously in a 2 × 2 matrix with a random arrangement. The participants selected objects by clicking on them with the mouse. When selected, the white frame of an object turned red. A second mouse click on the same object would deselect it, indicated by the frame's turning white again. The participants could select and deselect objects as often as they liked. They finished the trial by clicking on a button reading "none else."

In Phase 3, the participants evaluated four sets of four conditional inferences, preceded by a fifth set used as an example. Each set used the same conditional as a major premise; each object was used once for a set in a random order. Over the four sets, we realized a 2 × 2 design by crossing frequency of the antecedent with frequency of the consequent, in the same way as that for the selection task. The four inference tasks within each set were formed by adding one of the four possible minor premises, thus generating the four inference forms MP, AC, DA, and MT in random order. The conditional premise was printed in the top third of the screen, the minor premise below it, and the proposed conclusion on the bottom. The participants were asked to evaluate whether the conclusion followed with logical necessity from the two premises. They indicated their decision by pressing the left or the right arrow key. The assignment of answers to keys was visible on the screen.

In the next phase, the participants attempted to solve 16 syllogisms (following 1 practice trial). Each syllogism used for its three terms three feature values of the same object (e.g., "All 50¢ coins have a yellow border. Some coins with a yellow border have a star"). The feature visible from both sides was used as the middle term (i.e., B). Each syllogism in Table 6 was presented once with three frequent feature values as terms and once with three rare values. The objects were assigned to syllogisms at random, and the order of the 16 trials was also random.

The two premises of each syllogism were displayed below each other on the screen. The participants first decided whether something followed logically from the premises. If they decided that something followed, they were presented with a display of the eight possible conclusions (four quantifiers times two orders of the end terms), from which they selected the correct one by pressing the appropriate number key. The order of the conclusions in this display was counterbalanced over participants but was kept fixed over the 16 trials.

In the final phase, the participants were asked to estimate the proportions of the values of the 12 features in random order. For each feature dimension, they saw a scale from 0 to 100 on the screen, with the two feature values written at the two ends. An arrow pointed at the middle of the scale, and the participants could move the arrow with two keys. The current position of the arrow was always displayed as a numerical ratio (e.g., 40:60) below the scale. When ready, the participants entered their estimations by pressing the return key.

There was only one difference between the two experiments: In Experiment 2, the participants received an additional hint to "consider whether the previously learned frequencies are relevant for this task" as part of the instruction text preceding each reasoning task (we used a slightly different wording each time to avoid monotony). This should have encouraged the participants to take the frequency information into consideration without forcing them to use information that is normatively irrelevant for a deduction task.

RESULTS

Probability Learning

Figure 1 shows the probability that the participants guessed the more frequent feature value over successive bins of 20 trials in the learning phase of each experiment. The guessing behavior quickly approached the pattern expected for probability matching: Values were guessed at a proportion closely matching their true proportion (i.e., 90:10).

At the end of both experiments, most of the participants correctly estimated which feature value had been more frequent. An estimate was regarded as correct if the more frequent value was estimated as being more frequent.

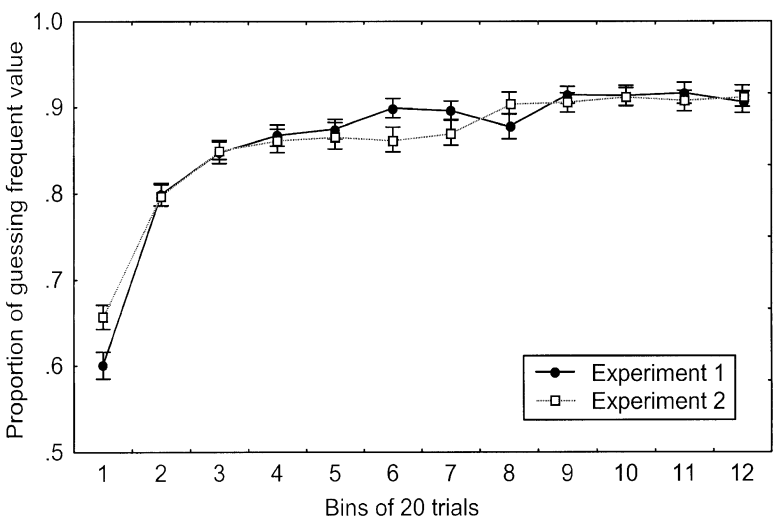

Figure 1. Means of the proportions of trials on which the participants guessed the frequent value during the learning phase in successive bins of 20 trials. Error bars represent one standard error.

Overall mean proportion correct was .90 in Experiment 1 and .92 in Experiment 2. Among correct estimates, the mean estimated proportions of the frequent value were .84 and .85, respectively. Among incorrect estimates, the means were .37 and .35, respectively. Many of the incorrect responses were 50:50 estimates. In Experiment 1, 34 participants selected the more frequent value as more frequent for all 12 features; 35 participants did this in Experiment 2. We conclude that most of the participants had learned the induced probability structure well and had memorized it by the end of the experiment.

Reasoning Tasks

Table 4 summarizes the selections of logical cases (corresponding to the cards in the original task) in the Wason selection tasks as a function of the probabilities of *p* and of *q*. The predictions of ODS were not confirmed for any of the four logical cases.

We computed contrasts capturing the predictions of the revised ODS theory for each card (see Table 1). These contrasts were computed on individual selection patterns within each condition, coding selection of an object as 1 and nonselection as 0. The theoretical range of the contrasts thus was –2 to 2; positive values indicate support for ODS. None of these contrasts deviated significantly from zero in Experiment 1. In Experiment 2, three contrasts deviated significantly from zero in the wrong direction, and the fourth was not significant. In both experiments, the pattern did not change when we limited analysis to the subset of 34 and 35 participants, respectively, who gave correct frequency estimates on all 12 questions at the end of the experiment.

The percentage of participants accepting each of the four conditional inferences in the four experimental conditions is summarized in Table 5. We computed contrasts capturing the predictions from the theory of Oaksford et al. (2000) for DA, AC, and MT (see Table 1). In Experiment 1, two of these contrasts were numerically negative and not significant; the AC contrast, however, was just significantly larger than zero in a one-tailed test ($p = .04$). In Experiment 2, none of the contrasts was significant, and the one approaching significance was negative. In the subsets of participants giving completely correct estimates at the end of the experiment, none of the contrasts approached significance.

The frequencies with which each quantifier was selected for each syllogism in the two experimental conditions (low vs. high frequency of the three terms) are shown in Table 6, together with the number of *no valid conclusion* responses. To test the hypothesis of PHM, we counted, for each participant, the number of positive quantifiers (*all* or *some*) and the number of negative quantifiers (*no*, *some . . . not*) chosen in each condition over the first six syllogisms in Table 6. As was argued above, PHM should predict that there are more positive and fewer negative quantifiers when the three terms are frequent than when they are rare. This was not the case in Experiment 1 [$t(67) = -0.66$ for the difference in the number of positive quantifiers between conditions, and $t(67) = -0.94$ for the number of negative quantifiers]. In Experiment 2, there were in fact more positive quantifiers selected when the terms were frequent (25%) than when they were rare [19%; $t(71) = -2.52$]. The difference for the negative quantifiers also went in the predicted direction (40% for frequent, 44% for rare terms), but the difference was not significant [$t(71) = 1.63$]. When we limited analysis to participants with completely correct estimates of proportions, none of the differences was significant in either experiment.

DISCUSSION

We tested predictions from the *probabilistic approach to human reasoning* advanced by Oaksford and Chater (2001) in three domains: the Wason selection task, conditional inferences, and syllogisms. The central empirical prediction that distinguishes this theory from others is that people's inferences should be affected by the probabilities of the categories the premises refer to. In our experiment, we used artificial materials in order to isolate inference processes from knowledge retrieval. We introduced probabilities by a probability-learning task, thus mimicking natural sampling (Gigerenzer & Hoffrage, 1995). The participants effectively learned and memorized these probabilities until the end of the experiment.

Nonetheless, we found little evidence that probabilities affected inferences in the two tasks involving conditionals. Only one contrast turned out to be significant in the expected direction in Experiment 1, and it was not replicated in Experiment 2. In addition, three contrasts in the Wason selection task turned out to be significant in the

Table 4
Percentages of Participants Selecting Each Logical Case in the Wason Selection Task

	Experiment 1				Experiment 2			
	LL	LH	HL	HH	LL	LH	HL	HH
Object *p*	68 (73)	63 (71)	75 (74)	69 (62)	64 (69)	46 (57)	69 (77)	71 (83)
Object *not-p*	10 (9)	16 (21)	13 (12)	15 (15)	22 (20)	25 (29)	10 (9)	06 (06)
Object *q*	40 (35)	44 (35)	44 (44)	37 (41)	50 (60)	53 (60)	38 (37)	44 (49)
Object *not-q*	16 (21)	21 (21)	12 (09)	13 (15)	21 (20)	17 (17)	29 (31)	14 (14)

Note—LL, LH, HL, and HH designate the experimental conditions formed by crossing probability of *p* (first letter: L, low; H, high) with probability of *q* (second letter). Values in parentheses are percentages for the subsamples of participants who answered all frequency estimation questions correctly ($n = 34$ and 35 for Experiments 1 and 2, respectively).

Table 5
Percentage of Participants Accepting Each Inference in the Conditional Reasoning Tasks

	Experiment 1				Experiment 2			
	LL	LH	HL	HH	LL	LH	HL	HH
MP	96 (97)	96 (97)	94 (94)	100 (100)	92 (89)	83 (83)	83 (77)	97 (100)
DA	43 (41)	41 (38)	28 (29)	43 (32)	53 (51)	47 (51)	55 (60)	58 (63)
AC	44 (41)	43 (47)	49 (47)	53 (56)	58 (63)	72 (71)	69 (69)	69 (66)
MT	62 (53)	57 (56)	66 (62)	65 (56)	59 (60)	60 (63)	66 (69)	70 (74)

Legend: See Table 1.

wrong direction in Experiment 2. The latter finding suggests that the instruction manipulation in Experiment 2 was successful. The participants took the frequency information into account, but not in the way predicted by ODS. Instead, it seems the instruction induced them to pay more attention to the more frequent cards, and this could have led them to select them preferentially (cf. Evans, 1996). Some, albeit weak, support was obtained for the predictions of PHM for syllogisms in Experiment 2.

Could the null results on most of our comparisons be due to a lack of power? For each of our experiments, the power to detect small effects ($f = .10$) was .20; for medium effects ($f = .25$), the power was .65; and for large effects ($f = .40$), it was .95 (computed by G*power; Erdfelder, Faul, & Buchner, 1996). Thus, there is a chance that we missed small or even medium-sized effects in some of our statistical tests. Most of the contrasts, however, failed the conventional significance level in both experiments. Moreover, support for Oaksford and Chater's (2001) theory requires that not just one or two but all the contrasts, at least within each task paradigm, be significant. The probability that we missed a whole set of consistently positive contrasts in both experiments is small, even when we assume small effect sizes.

Our findings are not inconsistent with previous results. The results on the Wason task are in full agreement with the experiments of Oberauer et al. (1999). Oaksford et al. (1999) found occasional support for ODS theory in four experiments with the Wason selection task. In each experiment, only some of the predicted effects were significant. Two of the experiments manipulated probability through the use of different thematic materials. Therefore, the observed effects could have been due to confounds with unknown characteristics of the materials. Thus, Oaksford et al.'s (1999) results provide only weak support for ODS. Pollard and Evans (1983) used a probability-learning task to induce frequency information prior to testing with the Wason selection task. Different from our experiments, they manipulated the frequency of conjunctions (i.e., truth table cases) and, thereby, the probability that the rule was

Table 6
Number of Participants Selecting Each Quantifier, As Well As "No Valid Conclusion," in the Conclusion for Each Syllogism, Broken Down by Condition

	All		Some		No		Some . . . Not		No Valid Conclusion	
Syllogism	L	H	L	H	L	H	L	H	L	H
All B are A,	1	5	3	2	35	34	2	5	27	22
No C is B	4	7	4	3	35	34	5	1	24	26
All B are A,	3	6	2	0	40	39	4	3	19	20
No B is C	10	12	7	6	31	34	4	6	20	14
No A is B,	5	8	1	2	32	36	2	1	28	21
All B are C	5	10	3	6	35	34	2	1	27	20
No B is A,	3	7	1	2	32	36	3	2	29	21
All B are C	7	11	9	3	33	29	1	3	21	25
Some A are B,	1	2	17	12	3	1	23	21	24	25
Some B are not C	2	7	11	13	2	2	24	17	33	33
Some A are not B,	0	3	22	17	3	2	10	14	33	32
Some B are C	5	6	15	24	3	1	17	11	30	30
Some A are B,	5	5	39	38	0	1	11	7	13	17
Some B are C	8	8	29	31	2	1	9	7	24	25
No A is B,	12	11	5	7	22	15	2	1	27	34
No B is C	13	13	11	10	17	20	6	7	24	22

Note—L, low frequency of the terms (A, B, and C); H, high frequency. First row, Experiment 1; second row, Experiment 2. Conclusions were statements relating the terms A and C by the quantifier denoted in the top row; conclusions using the terms in orders A–C and C–A are added together. $N = 68$ for Experiment 1; $N = 72$ for Experiment 2.

true. The finding that this manipulation affected the pattern of card selections is not in conflict with our results, and it does not support ODS (cf. Evans & Over, 1996b).

Oaksford et al. (2000) reported results in favor of their account of conditional reasoning. The procedures in all three experiments deviated from those of the standard deductive inference paradigm used here. Their first experiment introduced the conditional as the hypothesis of a protagonist, and the participants were not instructed to take it as a given premise. In the second experiment, the conditional was introduced as a completely redundant summary of the frequency information given. This might have induced the participants to reason from the frequencies directly, which would yield the predicted results. The third experiment introduced probabilities through thematic materials. Hence, it is again unclear to what degree the participants' responses rested on knowledge instead of reasoning and knowledge would bias responses toward the predictions (e.g., people would be less likely to infer that a person is a politician than to infer that a vegetable is eaten cooked, simply because the latter is more likely to be true in the real world). Thus, there is as yet no convincing evidence that probabilities affect people's conditional reasoning processes under standard conditions of deductive inference tasks. For the syllogism tasks, a direct test of PHM by manipulating the rarity of the terms has, to our knowledge, not been attempted before.

To conclude, our results suggest that probability manipulations through a natural sampling procedure have little effect on conditional inferences and on the Wason selection task. The effects we observed were either unsystematic or in the opposite direction of what Oaksford and Chater (2001) predicted. Our findings seriously question the assumption that experimental manipulations of probabilities of the categories involved in reasoning tasks affect the reasoning process in the way predicted by the probabilistic approach to human reasoning. This is not sufficient to refute the theory of Oaksford and Chater (2001), but it undermines the key assumption they used for linking the theory to empirical tests and, thereby, leaves it with little support.

REFERENCES

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

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavioral Research Methods, Instruments, & Computers*, **28**, 1-11.

Evans, J. S. B. T. (1996). Decide before you think: Relevance and reasoning in the selection task. *British Journal of Psychology*, **87**, 223-240.

Evans, J. S. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction.* Hove, U.K.: Erlbaum.

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

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

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

Handley, S., Feeney, A., & Harper, C. (2002). Alternative antecedents, probabilities, and the suppression of fallacies in Wason's selection task. *Quarterly Journal of Experimental Psychology*, **55A**, 799-818.

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

Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological problems of confirmation. *Thinking & Reasoning*, **2**, 1-31.

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

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

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

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

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

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

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

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135-151). Harmondsworth, U.K.: Penguin.

NOTE

1. Oaksford and Wakefield (2003) argued against this interpretation on the basis of an analysis of selections over the serial order of the 40 cards. The critical prediction distinguishing ODS from an avoidance-of-redundancy strategy is that the probability to select a *not-p* card should increase over the sequence when *not-p* is rare; the same should hold for *not-q* cards when they are rare. This trend was significant for *not-p* cards, but not for *not-q* cards. Thus, the support for ODS is still ambiguous.

(Manuscript received July 15, 2002;
revision accepted for publication June 30, 2003.)