Psychological Research (2001) 65: 128–144 © Springer-Verlag 2001

ORIGINAL ARTICLE

Magda Osman · Donald Laming

Misinterpretation of conditional statements in Wason's selection task

Received: 26 February 1999 / Accepted: 5 November 1999

Abstract Errors may be made on Wason's selection task because either (a) the rule to be tested is misunderstood, or (b) reasoning from that rule is inaccurate, or both. We report two experiments using the experimental paradigm introduced by Gebauer and Laming in which subjects are given six problems in succession. We use the subset of cards selected by each subject as (a) an indication of how the rule is understood and, when that selection is consistent throughout all six problems (so that we can infer a consistent understanding of the rule), as (b) a basis for evaluating the accuracy of the subject's reasoning according to three independent criteria. Experiment 1 adds an exactly parallel contextual version of the task to permit comparison between performances (by the same subjects) on the two versions. Experiment 2 repeats Exp. 1, but with negatives inserted in the conditional rule. Most subjects make a consistent selection of cards throughout all six problems, but typically appear to misunderstand the rule. This is so in both abstract and contextual tasks and replicates the finding by Gebauer and Laming. Most misunderstandings consisted of either (a) reading the simple conditional rule as a bi-conditional or (b) substituting "top/underneath" for "one side/other side". In Exp. 1 subjects seldom misevaluated the rule they appeared to be testing, but such "errors" of evaluation were common in Exp. 2. Negatives confuse the subjects and should not be used in any conditional application that matters. In Exp. 2 (but not 1) there was a significant correlation between interpretations of the two tasks. We provide an explanation of "matching bias" (it results from the confluence of the two common misunderstandings above) and comment on "mental models" which are, at present, unable to accommodate the variety of results we present here. We also relate our experimental paradigm to the conditional inference task and to truth tables.

M. Osman
Brunel University

D. Laming (✉)
Department of Experimental Psychology,
Downing Street, Cambridge, England CB2 3EB
e-mail: drjl@cus.cam.ac.uk
Tel.: (44) 1223 333565,
Fax: (44) 1223 333564

Introduction

We replicated the first experiment by Gebauer and Laming (1997) on Wason's (1966) Selection Task. In that experiment there were four cards, each with a letter on one side and a number on the other, showing, let us say *E*, *K*, *4*, & *7* (i.e. *V, C, e,* & *o,* introducing the notation – *V* for a vowel, *C* for a consonant, *e* for an even number, and *o* for an odd number – that we shall use here). Subjects were asked to determine by turning the cards over, but only as few cards as were necessary, whether it was true that: "If a card has a vowel on one side, then it has an even number on the other." Gebauer and Laming gave their subjects six successive such problems without any feedback from the experimenter. In some of those problems the rule held and in others it did not. We shall explain first why the six successive problems constituted an important innovation.

The correct answer to the problem is to turn over *V* and *o* and those two cards only. However, very few subjects find this correct answer. Johnson-Laird and Wason (1970a), reviewing four experiments with university students (Wason, 1968; Wason, 1969; Wason & Johnson-Laird, 1970), reported that only 4% of their subjects answered correctly. As a consequence, Wason's selection task has become the preferred vehicle for studying illogicalities in reasoning.

There are two categories of explanation one might advance for failing this task:

1. Subjects misunderstand the rule they are asked to test (different subjects misunderstanding in different ways), but thereafter respond in logical accord with the rule as they understand it.

2. Subjects do not reason logically.

As long as subjects are given only *one* trial, it is not possible to distinguish between these two broad categories of explanation; and as long as subjects are asked merely to *say* which cards need to be turned over, a second trial is not meaningful – subjects would simply repeat their first-trial response. (There have been studies in which the same subjects have been set several related, but different, problems – e.g., Bracewell & Hidi, 1974; Evans & Lynch, 1973; Griggs & Cox, 1983 – but the differences between the problems in those studies preclude the inferences we draw below.) However, if subjects are asked to physically turn cards over and determine whether the rule is true – sometimes it is, sometimes it is not – a meaningful succession of problems can be presented. Gebauer and Laming (1997) found that 44 out of 68 subjects (65%) in their two experiments performed with complete logical consistency with respect to some (mis)interpretation of the rule to be tested. A further 7 subjects showed just one change of mind.

We will show in what follows that there are *two* sources of error in a reasoning task. Responses might be incorrect because of:

1. a misperception of what the task consists of, or
2. a failure to reason accurately from that misperception,

or, of course, both. This prescription applies to any reasoning task, though we shall demonstrate it only with respect to Wason's selection task. A model for a reasoning task must therefore incorporate *two* stages to correspond to these two sources of error, and evaluation of such a model requires the experiment to identify how each individual subject perceives the task (in the selection task, how each individual subject interprets the rule).

Misunderstandings of the rule

Gebauer and Laming (1997, Table 4) noted two component misinterpretations of the rule. The rule, as set out above, is a simple conditional, but most subjects took it to be a bi-conditional, implying, in addition, "If a card has an even number on one side, then it has a vowel on the other." Given that interpretation, it is logical to examine all four cards. The other component misinterpretation was to read "top/underneath" for "one side/other side" ("If a card has a vowel on top, then it has an even number underneath."). Given that interpretation, it is logical to examine *V* only. When both component misinterpretations are combined, implying, in addition, "If a card has an even number on top, then it has a vowel underneath", both *V* and *e* are examined.

The suggestion that subjects might understand Wason's rule in different ways has quite a long history. Smalley (1974) considered the possibility of different interpretations of conditional statements (of which bi-conditionality might be one) by different subjects, and Johnson-Laird and Wason (1970a, p. 141) noted that "the reversibility of the cards is not always recognized." More recently, Johnson-Laird and Byrne (1991, p. 80) and Johnson-Laird (1995, p. 136) have envisaged, in the context of the "mental models" theory, that subjects might interpret the rule as either a simple conditional or a bi-conditional. Bi-conditional truth functions have also been suggested by Taplin (1971), and Taplin and Staudenmayer (1973) remarked on "different interferences by the subject about how the [conditional] sentence is supposed to be used in the experimental task brought about by relatively small differences in the procedure employed." But these last two studies were concerned, not with Wason's selection task, but with the evaluation of conditional arguments. What no one had done, until Gebauer and Laming (1997), was to submit the matter to experimental examination.

"Logicality"

Gebauer and Laming (1997) interpreted their findings to mean that when subjects select the wrong cards, most often this is because they misunderstand the rule that is to be tested and select, instead, those cards which are logically appropriate to testing the rule as they understand it to be. Stenning and van Lambalgen (1999) have questioned that interpretation, suggesting that subjects may be doing no more than *selecting* cards consistently. To resolve the issue of what "logicality" actually comprises in relation to Wason's selection task, suppose that, historically, subjects had always got the task correct, that is, had always pointed out the *V* and the *o* as the cards needing to be examined. No one would then have suggested that such responses were anything other than logical. Gebauer and Laming have, as it were, fitted a model with one free parameter to 44 of their subjects. That model says that subjects respond logically, and the free parameter is the rule to which they are responding. Except that previous investigators have taken it for granted that the errors were *illogical* (explanation 2 above), there would be no further question. However, with the enhanced experimental paradigm introduced by Gebauer and Laming, we can study the question of logical performance in more detail.

Let us suppose, for the sake of exposition, that the rule is correctly understood. The instructions we give our subjects then intend:

(a) that they should turn over *V* and *o* and those two cards only;
(b) that they should correctly announce whether the rule holds or not; and
(c) that when the rule is false, they should make that announcement as soon as possible (if the *V* has an *o* on the underside, there is then no need to look at the *o*) and then stop.

Even amongst those subjects who consistently select V and o, these instructions might be realised less than completely in three respects:

1. Subjects might not always evaluate the rule correctly.
2. Subjects might turn over additional cards (C or e) and then turn them back again to indicate that those cards were not needed.
3. Subjects might go on to examine the o, even when the V is discovered to have an o underneath and the rule can already be known not to hold.

All three categories would be deemed correct within the traditional paradigm in which subjects are merely asked which cards need to be turned over, because these elements of performance are not then examined. However, Category 1 is less than logical; Category 2 includes an error, but that error is corrected without feedback (the underside of C or e tells the subject nothing about whether the rule holds); and Category 3 is arguably an easier strategy (turn over a fixed set of relevant cards and then make one decision) than the procedure intended in the instructions. We shall examine these three elements of performance which the traditional paradigm overlooks.

Previous work

At this point in an introduction, it is conventional to review existing theory to provide a context for the questions which are to be put to experimental test. The common view on which existing theory is based seems to be that, proceeding from the rule set out above, it requires an exercise of reason to identify V and o as the cards which need to be examined. Different theories have proposed different defects in that reasoning process. However, unless there is some intermediate observation between the presentation of the rule and the selection of the cards, that defective reasoning is not distinct from a simple misunderstanding, except that the notion of "misunderstanding" might also incorporate influences of context and of prior experience and training. If some simple formula had been discovered which accounted for virtually all of the errors on Wason's task, the idea of a defect in reasoning would have been credible. However, it happens that the card selections by different subjects are very varied (a wide variety of *different* defects in reasoning would be required), and for a majority of subjects the selections are entirely consistent over successive problems (the different defects in reasoning would need to be *systematic* within the individual subject). In addition, it is manifest that context assists subjects towards a correct answer. Although the notions of "defect of reasoning" and "misunderstanding" are not entirely distinct, it makes good sense to delineate an initial stage of potential misunderstanding of the task (source 1 above) and proceed from there.

We know from the work of Gebauer & Laming (1997) that a majority of subjects display a consistent misunderstanding over six successive problems. It does not follow, however, that the remainder suffer some logical defect; the absence of a consistent selection of cards might simply indicate an unstable state of mind or even no understanding at all. Previous theories have all addressed a presumed illogicality in the subjects' reasoning (Explanation 2 above) which has yet to be demonstrated. It is far from clear that existing theory has any relevance to the conduct of this investigation.

Work on Wason's selection task has hitherto been characterised by an extreme poverty of experimental finding. Most often, theory has looked at no more than the relative frequencies with which the four different cards are selected by subjects who attempt the problem once only (e.g., Kirby, 1994; Oaksford & Chater, 1994; Sperber, Cara, & Girotto, 1995). Moreover, no one has attempted to model the *absolute values* of those frequencies – only their rank order has been subject to enquiry. The original version of the task (with the rule formulated as above) has nearly always yielded the ranking $\langle V, e, o, C\rangle$ (see Oaksford & Chater, 1994, Table 2), while contextual versions of the task have sometimes inverted the order of the inner two elements (Cheng & Holyoak, 1985; Cosmides, 1989; Griggs & Cox, 1982). There are two reasons for this lack of progress:

1. As we have already explained, most existing theories, and previous analyses of data, take it for granted that subjects understand the rule they are to test and propose some systematic defect of reasoning to generate the spectrum of incorrect selections observed. This would appear to be addressing the wrong question. There has not, previous to Gebauer and Laming (1997), been any systematic exploration of what different misinterpretations subjects might make of the rule in the selection task.
2. Repeated failure to discover a systematic defect of reasoning that is also convincing has led to a proliferation of theoretical ideas. However, since there are so few empirical findings to be accommodated, most of those ideas must be empirically equivalent. Most precisely, most of the theoretical content is invention, neither right nor wrong, but speculation which, to our way of thinking, runs too far beyond the scope of existing experimental findings.

A theory, any theory, is only as good as the data on which it is based, and the existing corpus of results from Wason's selection task is inadequate for this purpose. What is urgently required is not theoretical solutions, but a wider spectrum of experimental findings – not more of the same data, but different kinds of data – which will indicate what kind of theory should be entertained. That motivates this present study. We replicated the first experiment by Gebauer and Laming (1997), with some minor improvements in design. We added an exactly parallel contextual task which enabled a comparison between performance on the abstract and on the contextual task by individual subjects. We then

repeated that comparison in a second experiment, but with negatives inserted into the rule in order to discover how subjects interpret negative conditionals. Throughout these procedures we asked our subjects (a) to say whether the rule holds or not, (b) to turn back again any cards which did not need to be examined, and (c) to stop as soon as they are able to say whether the rule was true or false. We then analysed our subjects' performances on two distinct levels; (a) the particular selection of cards which tells us how that subject has understood the rule and (b) the logicality of the reasoning with respect to that (mis)understanding. We also looked for correlations in performance both between these two levels and between the abstract and contextual tasks. To the extent that our data bear on existing theory, this is a matter of accident, rather than design, but on that basis we are able to comment in the Discussion on "Matching bias" and on "Mental models."

Experiment 1

So far, only the experiments by Gebauer and Laming (1997) have tested the idea that most subjects are entirely consistent in their selection of cards. In view of the theoretical implications of this result, it needs to be replicated. It is also worth replication with a contextual version of Wason's task, and we shall take this opportunity to compare individual performances on abstract and contextual tasks. Different subjects appear to understand the same rule in different ways, and we conjecture that the differences are idiosyncratic. If this is so, the same idiosyncrasies might be apparent in both abstract and contextual tasks. We therefore looked for some correlation between individual subjects' performances. However, we also conjecture that the way in which the rule is understood is constrained by the context in which it is presented – hence the greater proportion of correct solutions in contextual tasks. The correlation between individual subjects' performances may be attenuated for this reason. Finally, we shall examine the nature of logical performance in some detail.

Methods

Stimuli. There were two versions of Wason's task, (a) the original abstract version with a vowel (*V*) or consonant (*C*) on one side of the card and an even (*e*) or odd (*o*) number on the reverse, and (b) a contextual version using photographs of cars at traffic lights. Each photograph showed a traffic light at red (*R*) or green (*G*) and a car either stopped (*s*) or proceeding (*p*) beyond the stop line. The photographs were cut in two and the two halves stuck back to back, so that one side showed the traffic light and the other the car. There were six different sets of cards for each version of the task.

Instructions. The instructions for the two tasks were made as similar as possible to enhance comparability between them:

A. Abstract task. "You will be shown six sets of four cards. There will be a letter on one side of the card and a number on the other side. Please read the following carefully:
– If there is a vowel on one side of the card, then there is an even number on the other.
– Turn over as many cards as you need to discover whether this rule holds; and let me know as soon as you have discovered whether it holds or not. Do not turn over any more cards than you need. If you turn over a card and later decide you didn't need to do so, turn it back again. Please ask questions if there is anything you are unsure of."

B. Contextual task. "You will be shown six sets of four cards. There will be a traffic light on one side of the card and a car on the other side; they are two sides of the same photograph. Please read the following carefully:
– If there is a red traffic light on one side of the card, then the car on the other side is stopped behind the stop line.
– Turn over as many cards as you need to discover whether this rule holds; and let me know as soon as you have discovered whether it holds or not. Do not turn over any more cards than you need. If you turn over a card and later decide you didn't need to do so, turn it back again. Please ask questions if there is anything you are unsure of."

Design. Two sets of cards in each task (presented first and last) were correctly matched (*V* paired with *e* and *C* with *o*; *R* paired with *s* and *G* with *p*). The other four sets had one mismatch each, the mismatch occurring on the underside of a different card (*V*, *C*, *e*, *o*/*R*, *G*, *s*, *p*) in each one of the four. Comparisons between the first five problems enable us to distinguish as wide a variety of different interpretations of the rule (different selections of cards and corresponding evaluations of truth) as possible within that space of observation (Gebauer & Laming, 1977, distinguished only four different interpretations, though, to be sure, the most common four), and comparison of Problems 1 and 6 enable us to discover whether some particular interpretation had been retained throughout.

There were four layers of randomisation, intended to obviate effects of the order in which individual problems, cards, and tasks were presented:

1. Each subject attempted all six problems on each task. Problems 1 and 6 always had all four cards matching, the two matching sets being assigned at random to first and last positions. Problem 2–5 were ordered differently for different subjects, all possible permutations (24) being used equally.
2. The spatial arrangement of the cards as they were laid out for each subject was also varied. First, each set of cards was arbitrarily permuted from the standard order (*V*, *C*, *e*, *o*/*R*, *G*, *s*, *p*). Those arbitrarily permuted orders were further permuted before presentation, the same further permutation being applied to all six problems, but a different permutation for each subject.
3. The same permutations, 1 and 2, were applied to both abstract and contextual tasks, so that, while each subject had Problems 2–5 and the spatial layout of the sets of cards arranged differently, those different arrangements were exactly parallel for each subject with respect to the two tasks.
4. Finally, subjects were randomly assigned to receive one task or the other first.

Procedure. Subjects were first screened for red/green colour blindness using the Ishihara (1967) test. (Subjects who did not pass the Ishihara test were still tested, but their data were not included in the analysis.) Each subject was tested individually. They were given the instructions appropriate to the task they were to attempt first. When they indicated that they understood the instructions, they were given the six problems in order, being allowed to take as long as they wished over each. The rule to be tested was left exposed meanwhile. When a subject had examined the cards for one problem and announced whether the rule held or not, the next

problem was presented. Each card turned over was recorded, as well as the order in which the cards were turned, and those cards which were turned over and turned back again. The point at which the subjects announced whether the rule was true or false and any spontaneous comments were also recorded. When subjects had finished the first task they went on to the second. Finally, the background and aims of the experiment were explained.

Subjects. A total of 48 subjects were tested, 28 men and 20 women. They were all experienced car drivers, mostly professionals working at a local general hospital in Weston-super-Mare; two were in full-time education. Their ages ranged from 28–65. They had no prior knowledge of the task.

Results

In the original version of Wason's task, subjects were merely asked to *say* which cards they would need to examine, without turning anything. Each subject's response was therefore some combination of *V*, *C*, *e*, and *o* (or *R*, *G*, *s* & *p*). We first converted each observed sequence of card-turnings into such an equivalent combination, but subject to the following proviso (which exemplifies what we mean by "consistency"). If a subject has in mind to examine *V* and *e*, but the first of these turns out to have an odd number underneath, there is no need to examine the *e*: the rule is demonstrably false (and the subject is explicitly instructed to say so at this point and to stop). The subject's response is nevertheless compatible with the selection of *V* and *e*. It is also compatible with the selection of *V* alone, but, if the pattern of response on the preceding (or following) problem is also a $\langle V, e\rangle$ combination (using $\langle \ldots \rangle$ to indicate specifically a combination of cards selected by a subject), it is scored as $\langle V, e\rangle$. (The design of the experiment precludes both the preceding and the following pattern being compatible with both $\langle V, e\rangle$ and with $\langle V\rangle$ alone, so there is little confusion possible on this matter.) Subject to this proviso, many subjects selected the same combination on all six problems and are classified in Table 1 according to that combination.[1] Other subjects changed their selection once during the task. Those few who appeared to select more than two different combinations are categorised as "inconsistent."

Statistical methods. All the hypotheses tested concern associations between the row and column classifications in contingency tables. To test such a hypothesis, we routinely use Pearson's X^2, without Yates' correction and without restriction on expected frequencies, on the basis of the Monte Carlo studies of small samples by Larntz (1978) and Bradley, Bradley, McGrath, and Cutcomb (1979). When, however, we need to decompose a contingency table into statistically independent subcomponents, we use the log-linear G^2 (Agresti, 1984; Norusis, 1985) instead.

[1]A copy of the raw data may be had on application to either author.

Table 1 Classification of performances in Exp. 1 and comparison with Gebauer & Laming (1997, Exp. 1)

Combination		Contextual task	Abstract task	Gebauer & Laming, Exp. 1
Contextual	Abstract			
R	*V*	6	7	4
s	*e*	1	1	0
p	*o*	5	0	0
R,*G*	*V*, *C*	1	0	0
R, *s*	*V*, *e*	5	17	14
R, *p*	*V*, *o*	9	1	1
G, *s*	*C*, *e*	0	1	0
G, *p*	*C*, *o*	2	0	0
R, *G*, *s*	*V*, *C*, *e*	1	0	0
R, *G*, *p*	*V*, *C*, *o*	0	2	0
All 4 cards		13	6	3
1 change		3	10	4
Inconsistent		2	3	4
Total		48	48	30

Effect of task order. There were no significant differences on either task between the performances of those subjects who did that task first and those who did it second (abstract task: $X^2 = 11.815$ with 8 *df*, $p \approx 0.1$; contextual task: $X^2 = 16.349$ with 10 *df*, $p > 0.05$). We therefore aggregated both orders of administration in all subsequent analyses.

Consistency of card selection. Counting each subject twice (once for the abstract task, once for the contextual), 78 subjects (out of 96) were entirely consistent in their card selections, 13 changed their selection once, and only 5 were otherwise inconsistent. These proportions do not vary significantly between the two tasks, nor in comparison with Gebauer and Laming (1997, Exp.1; $G^2 = 7.019$ with 4 *df*, after aggregating the first 11 rows in Table 1). Any difference between the tasks resides only in the combination of cards chosen by those subjects who responded consistently throughout (i.e., in rows 1–11).

Comparison of abstract and contextual tasks. Looking, now, only at rows 1–11 in Table 1, there is no significant difference between the combinations of cards consistently selected in the abstract task here and in Gebauer and Laming (1997, Exp.1; $G^2 = 4.693$ with 10 *df*). However, the difference between the contextual task and these two abstract tasks taken together is highly significant ($G^2 = 42.606$ with 10 *df*, $p < 0.001$). This difference is chiefly due to the combinations $\langle R, s/V, e\rangle$, $\langle R, p/V, o\rangle$, $\langle p/o\rangle$ and ⟨all 4 cards⟩ (which contain most of the instances). The remaining categories do not differ significantly ($G^2 = 10.586$ with 6 *df*, $p \approx 0.1$).

"Logicality". Of the 78 subjects who selected cards consistently throughout, 4 (2 in each version of the task) made one or two errors of evaluation each.

Out of the 96 subjects, 23 turned one or more cards back again as not needing to be examined. Among those 23, the apparent understanding of the rule differed

between the abstract and the contextual tasks ($G^2 = 11.922$ with 4 *df*, $p < 0.02$) with the categories $\langle V, e \rangle$ and "1 change" prominent in the abstract task, but not in the contextual. Of the 30 (total) occurrences, 20 consisted of examining all four cards before turning some of them back again.

When a mismatch is discovered on the underside of a card, it is possible to declare the rule to be false straightaway without looking at any other potentially relevant cards. This can occur only on the middle four problems, and whether it does occur depends on the order in which the different cards are examined. On the abstract task there were 18 occasions on which a subject might have stopped short, and on the contextual task 45; but on *all 63 occasions* the subjects carried on.

Most of these numbers are too small for further analysis here, but we shall re-examine these data later when comparison can be made with the corresponding data from Exp. 2.

Correlation between individual performances on the two tasks. Examination of the contingency table "selection on abstract task" × "selection on contextual task," suitably reduced to aggregate infrequent response patterns, failed to reveal any significant association between response combinations on the abstract and on the contextual task ($X^2 = 28.114$ with 30 *df*). Restriction of this analysis to those subjects who made consistent card selections throughout modified the statistic to $X^2 = 17.211$ with 15 *df*, which is still not significant.

The numbers of errors of evaluation and of continuations, notwithstanding that the underside of a card already turned over showed the rule to be false, are too few to exhibit any correlation between the two tasks. However, it is worth remarking that the correlation for turning cards back again as not needing to be examined was exactly zero ($X^2 = 0$).

Commentary

We found no differences consequent on the order in which the two tasks were administered. This confirms the work of Griggs and Cox (1982), who tested for transfer from a contextual task to the abstract. This suggests to us that most subjects do not see the two tasks as variants of a common logical problem, but as unrelated. However, one subject did comment, "So, is there any difference between the two questions I have done, did I answer the same way?" That subject switched from $\langle V, e \rangle$ to ⟨all 4 cards⟩ during the abstract task and then continued selecting ⟨all 4 cards⟩ on the contextual task.

The arrangement of the mismatches on Problems 2–5 was designed to distinguish as wide a variety of different interpretations (that is, card selections combined with evaluations of the truth of the rule) as possible. However, it is possible that subjects do not have a precise conception of what they understand the rule to be. For example, Wason and Evans (1975) found their subjects were unable to justify their selections and could not articulate exactly what they took the rule to mean. Some of our subjects also seemed less than certain. Of six subjects who examined all four cards on the abstract task, four spontaneously said something like, "I have to make sure they are all right, if the rule holds. I can't be absolutely sure unless they are all turned over."

It may alternatively be that a subject's understanding crystallises in the course of attempting several successive problems. However, this much is noteworthy: selection of most of the 15 different combinations of cards cannot be summarised by any simple rule, but the most frequently selected combinations can. These are set out in Table 2; they are the ones specifically trapped by Gebauer and Laming (1997). The first two interpretations distinguish between the *top* of the card (which the subject can see) and the *bottom* (which is not visible); the second two interpretations make no such distinction. The second and fourth interpretations are bi-conditional (an even number is always paired with a vowel); the first and third are not. The third interpretation is correct.

It should be noted that each of the simple rules in Table 2 identifies a unique subset of cards as needing to be examined; conversely, each of the other 11 subsets of cards corresponds to some other rule (which does not admit any simple formulation). Those other 11 subsets are sometimes selected and selected consistently (see Table 1), and an open questions is: What was in those subjects' minds at the time? We have no answer to suggest here, but a practical way to find out would be to prepare a set of supplementary problems for presentation when such an atypical understanding of the rule is discovered.

The contextual version of Wason's task, using pictures of cars at traffic lights, was chosen to be sensitive to prior experience, dispensing with the sometimes elaborate scenario that is commonly used to set the scene in contextual tasks (Cheng & Holyoak, 1985; Cosmides, 1989; Gigerenzer & Hug, 1992). Although we did not interrogate or debrief our subjects after the experiment, some of them made spontaneous comments which indicated that the influence of prior experience contributed in the way we expected. During the contextual task four subjects said something about drivers breaking the law, such as, "I have to do this to see if they are breaking the law."

This influence of prior experience led, relative to the abstract task, to an increased proportion of correct selections $\langle R, p/V, o \rangle$, as is commonly found in contextual tasks (Cheng & Holyoak, 1985; Cosmides, 1989; Girotto, Mazzocco, & Tasso, 1997; Sperber, Cara, & Girotto, 1995) and also of selections of all four cards. This latter pattern of response may be specific to our road-traffic task. If a traffic light shows green, one ought, as a matter of courtesy, to move away promptly, even though one would have to pause behind the stop line for a substantial time before one could be charged with obstruction. These two patterns of response increased at the expense of the two cards mentioned in the rule $\langle R, s/V, e \rangle$, which is the most common pattern in the abstract task.

Table 2 The most common card selections and their interpretations

Card selection	Interpretations of the rule	
	Abstract task	Contextual task
V/*R*	If there is a vowel on top, there is an even number underneath.	If there is a red traffic light on top, then the car underneath is stopped behind the stop line.
V, *e*/*R*, *s*	If there is a vowel on top, there is an even number underneath, and if there is an even number on top, then there is a vowel underneath.	If there is a red traffic light on top, then the car underneath is stopped behind the stop line, and if the car on top is stopped behind the stop line, there is a red traffic light underneath.
V, *o*/*R*, *p*	If there is vowel on one side of the card, then there is an even number on the other side.	If there is a red traffic light on one side of the card, then the car on the other side is stopped behind the stop line.
All 4 cards	If there is vowel on one side of the card, then there is a even number on the other side, and if there is an even number on one side, then there is a vowel on the other.	If there is a red traffic light on one side of the card, then the car on the other side is stopped behind the stop line, and if the car on one side is stopped behind the stop line, there is a red traffic light on the other side of the card.

Experiment 2

Experiment 2 repeats the design of Exp. 1 exactly except for the use of negatives in the rules to be tested. We thereby combined the six-problem sequence of Gebauer and Laming (1997) with negatives in the expression of the rule (the self-same rule) to see if the pattern of consistent responding continues to hold. We also replicated the comparison of different combinations of negatives in both abstract and contextual tasks by Griggs and Cox (1983, Exp. 2), but using our more elaborate experimental paradigm. Griggs and Cox chose conditional statements that reflected real-world situations with the idea that such a correspondence would facilitate correct reasoning, for example, ''If a person is drinking beer, then the person must not be under 19.'' Our use of pictures of cars at traffic lights and experienced drivers as subjects continues that practice.

Methods

Stimuli, design, and procedure. These were exactly the same as in Exp. 1 and need no further comment.

Instructions. These were also the same except for the insertion of negatives in the rules to be tested. We compared three different placements of negatives (as follows) and use the data from Experiment 1 as the 'no negative' control.

A. Abstract task.

1. ''If there is a vowel on one side of the card, then there is **not**[2] an odd number on the other.''
2. If there is **not** a consonant on one side of the card, then there is an even number on the other.''
3. ''If there is **not** a consonant on one side of the card, then there is **not** an odd number on the other.''

B. Contextual task.

1. ''If there is a red traffic light showing on one side of the card, the car on the other side has **not** proceeded over the stop line.''
2. ''If there is **not** a green traffic light showing on one side of the card, the car on the other side has stopped behind the line.''
3. ''If there is **not** a green traffic light on one side of the card, the car on the other side has **not** proceeded over the stop line.''

Subjects. A total of 72 subjects were tested, 41 men and 31 women, a different 24 subjects with each placement of negatives (1, 2, & 3). They were all experienced car drivers and were recruited through the South Cambridge Neighbourhood Watch. Their ages ranged from 28–82. None of them had any prior experience of the selection task, though 4 had had some training in logic.

Results

The raw data were recorded, converted to equivalent combinations of cards, and categorised exactly as described for Exp. 1 to give the classified data shown in Table 3.[3] The methods of statistical analysis are also the same. Since Exp. 1 failed to show any difference in performance dependent on the order in which the two tasks were administered, this factor was ignored.

Consistency of card selection. Counting each subject twice (again), 87 subjects (out of 144) were entirely consistent in their selection of cards, 15 changed their selection once, and 42 were otherwise inconsistent. These proportions do not vary significantly between the

[2] The negatives are set in bold here for the benefit of the reader; they were not highlighted in the instructions shown to the subjects.

[3] A copy of the raw data may be had on application to either author.

Table 3 Frequencies of card selections in Exp. 2

Combination		Abstract rules			Contextual rules		
Abstract	Contextual	If V, then $\neg o$	If $\neg C$, then e	If $\neg C$, then $\neg o$	If R, then $\neg p$	If $\neg G$, then s	If $\neg G$, then $\neg p$
V	R	5	1	0	4	1	0
C	G	0	1	0	0	0	0
V, C	R, G	3	0	1	0	0	1
V, e	R, s	1	3	3	2	0	2
V, o	R, p	1	1	2	7	4	2
C, e	G, s	0	1	1	0	1	0
C, o	G, p	0	0	0	0	1	2
e, o	s, p	0	0	1	0	0	0
V, C, e	R, G, s	0	0	1	0	0	0
V, C, o	R, G, p	0	0	1	2	0	3
V, e, o	R, s, p	0	0	1	1	0	1
All 4 cards		4	6	4	2	6	3
1 change		2	3	3	0	3	4
Inconsistent		8	8	6	6	8	6
Total		24	24	24	24	24	24

two tasks, nor between the three differently negated rules ($G^2 = 8.167$ with 10 *df*, aggregating the first 12 rows in Table 3). Any difference between the tasks resides only in the combinations of cards chosen by those subjects who responded consistently throughout (i.e., in rows 1–12).

Comparison of abstract and contextual tasks. Looking, now, only at rows 1–12 in Table 3, there is no significant difference between the combinations of cards consistently selected under the three different rules in the abstract task ($G^2 = 26.873$ with 22 *df*), nor in the contextual task ($G^2 = 26.621$ with 22 *df*), but the two tasks do differ from each other ($G^2 = 20.030$ with 11 *df*, $p < 0.05$).

"Logicality". Table 4 divides those subjects who made a consistent selection of cards throughout according to whether they were also always correct in their evaluation of the rule or were sometimes wrong. There is no significant difference in these numbers between the abstract and contextual tasks ($G^2 = 1.689$ with 1 *df*), but a comparison between the rules (pairing rules 1, 2, & 3 according to the location of the negative) gives $G^2 = 10.319$ with 2 *df*, which is significant at 0.01. Decomposing this statistic, there is no difference between the first value (If V, then $\neg o$ / If R, then $\neg p$) and the second (If $\neg C$, then e/If $\neg G$, then s), with one negative each, in the proportion of subjects who are sometimes wrong ($G^2 = 1.830$ with 1 *df*), but the third rule (If $\neg C$, then $\neg o$ / If $\neg G$, then $\neg p$) with two negatives shows a greater proportion of subjects making errors ($G^2 = 8.490$ with 1 *df*, $p < 0.01$).

In each version of the task, 19 subjects turned one or more cards back again as not needing to be examined. There was no significant difference between the abstract and contextual tasks in the apparent understanding of the rule by these subjects ($G^2 = 6.979$ with 5 *df*), most of whom (21 out of 38) were classified as "inconsistent." Of the 58 (total) occurrences, 25 consisted of examining all four cards before turning some of them back, and 23 consisted of examing three cards beforehand.

The 86 subjects who selected cards consistently throughout encountered, between them, 46 occasions when they might have declared the rule to be false before examining all of the potentially relevant cards. On 31 of those 46 occasions they continued with their examination; on 15 occasions they stopped.

Correlation between individual performances on the two tasks. Since there was no significant difference found (above) between the combinations of cards selected under the three different rules on both the abstract and on the contextual task, the data from the three rules were aggregated for the purpose of enquiring whether there was any correlation between individual performances on the two tasks. This gave a total of 72 entries in the contingency table "selection on abstract task" × "selection on contextual task." After suitable reduction to aggregate infrequent response patterns, the association between response combinations proved highly significant ($X^2 = 109.806$ with 49 *df*, $p < 0.001$). Restriction of the analysis to those subjects who responded consistently throughout gave $X^2 = 68.488$ with 25 *df*, which is again significant at 0.001. Visual examination of the contingency table shows that subjects tended to choose similar combinations of cards on the two tasks; 10 subjects (out of 72) were inconsistent both times and, amongst the 29 subjects consistent on both tasks, the combinations $\langle V/R\rangle$, $\langle V, e/R, s\rangle$, $\langle V, o/R, p\rangle$, $\langle C, e/G, s\rangle$, and ⟨all 4 cards⟩ were chosen 2, 3, 3, 1, and 8 times, respectively.

Looking now at the three criteria of "logicality" as they apply to those subjects who made consistent selections of cards on both tasks, there was no significant correlation of errors in evaluation ($X^2 = 2.535$ with 1 *df*), nor in turning cards back again as not needing to be examined ($X^2 = 0.386$ with 1 *df*). However, of four subjects who had the opportunity of stopping early on

Table 4 Evaluation of the rule in Exp. 2 by those subjects who made consistent selections throughout

Pattern of evaluation	Abstract rules			Contextual rules		
	If V, then $\neg o$	If $\neg C$, then e	If $\neg C$, then $\neg o$	If R, then $\neg p$	If $\neg G$, then s	If $\neg G$, then $\neg p$
Always correct	6	5	2	12	5	3
1 or more errors	8	8	13	6	8	11
Total	14	13	15	18	13	14

both tasks and declaring the rule to be false, three stopped on both tasks and one continued on both. This sample of size 4 is actually just significant at 0.05 ($X^2 = 4.0$ with 1 *df*).

Comparison with Experiment 1

Experiment 2 employed exactly the same stimuli, instructions, design, and procedure as Exp. 1. Moreover, the rule to be tested was also the same, both in Exp. 1 and in all three rule-conditions of Exp. 2; the only difference was the way in which the common rule was formulated. Within each task (abstract and contextual) in Exp. 2, the combinations of cards consistently selected did not differ between the different formulations and, for each formulation (different placement of negatives), the proportion of consistent card selections incorrectly evaluated differed relatively little in relation to the comparisons yet to come. Thus as a matter of convenience, Table 5 aggregates the data from the three rules in Exp. 2 to exhibit an important difference with respect to Exp. 1.

Consistency of card selection

Combining rows 1 and 2 (subjects who selected cards consistently) in Table 5, there remains a significant difference between the two experiments (abstract task, $G^2 = 12.338$ with 2 *df*, $p < 0.01$; contextual task, $G^2 = 13.953$ with 2 *df*, $p < 0.001$). To put the matter succinctly, negatives in the rule make the task harder.

Evaluation of the truth of the rule

Looking now at the difference between rows 1 and 2, many more of those subjects in Exp. 2 who selected cards consistently failed to evaluate the rule (as they appeared to have understood it) correctly. There is no difference between the abstract and contextual tasks within either experiment (rows 1 & 2 only: Exp. 1, $G^2 = 0.045$ with 1 *df*; Exp. 2, $G^2 = 1.382$ with 1 *df*), but a very obvious difference between experiments ($G^2 = 68.011$ with 1 *df*, $p < 0.001$).

Card selections

There are minor differences between the combinations of cards selected in Exps. 1 and 2 (abstract task: $G^2 = 22.111$ with 11 *df*; contextual task: $G^2 = 22.051$ with 11 *df*, both significant at $p < 0.05$). In the abstract task the predominant combination in Exp. 1 is $\langle V, e \rangle$; in Exp. 2 this is replaced by ⟨all 4 cards⟩. In the contextual task there is less difference, with $\langle R, p \rangle$ (which is the correct selection under all four rules) and ⟨all 4 cards⟩ both being common.

Turning cards back again

About the same proportion of subjects in each experiment turned cards back as not needing to be examined (Exp. 1, 23/96; Exp. 2, 38/144), but the apparent understanding of the rule by these subjects differed ($G^2 = 30.113$ with 6 *df*, $p < 0.001$). In Exp. 2 most of these subjects showed no consistent selection of cards at all.

Continuing after the rule could be known to be false

This was relatively less common in Exp. 2 (Exp. 1, 30/30; Exp. 2, 15/23), but examination of the apparent understandings of the rule by these subjects discovered no systematic difference between the experiments ($G^2 = 28.638$ with 28 *df*).

Table 5 Evaluation of the rule in Exps. 1 and 2

Pattern of selection and evaluation	Experiment 1		Experiment 2	
	Abstract task	Contextual task	Abstract task	Contextual task
Subjects making consistent selections throughout:				
Correct evaluation of rule	33	41	13	20
Incorrect evaluation of rule	2	2	29	25
Inconsistent subjects:				
1 change	10	3	8	7
Inconsistent	3	2	22	20
Total	48	48	72	72

Evolution of "logicality"

There is no evidence that conformity to the three criteria of "logicality" evolves over the course of the six problems presented in our experiments. Indeed, aggregating the data from both experiments, the frequencies of logical errors were remarkably uniform (false evaluations: $G^2 = 5.279$ with 5 *df*, mean = 21.5; turning cards back: $G^2 = 8.186$ with 5 *df*, mean = 14.33; continuing after the rule could be known to be false; $G^2 = 0.383$ with 3 *df*, mean = 23.5).

Commentary

We found no differences in the consistency of card selection in Exp. 2, either between the three rules or between the two tasks, abstract and contextual. Differences between the tasks did show in the form of different combinations of cards selected, but again there were no significant differences between the rules within each task. This might reasonably have been expected, since the three rules all had the same meaning – merely differently formulated with different placement of negatives. A difference between the three rules, but not now between the tasks, did, however, appear in the proportions of subjects who, though selecting cards with complete consistency, were not always correct in evaluating the truth of the rule as they appeared to understand it. For example, Problems 1 and 6 always had all four cards with top and bottom corresponding, so a selection which revealed Problem 1 to be "True" must necessarily have the same outcome on Problem 6. Not every subject was consistent in this way. Another subject might consistently examine all four cards every time and always say "False," whatever the correspondence between top and bottom. The rule with two negatives led, not surprisingly, to the greatest confusion of the three about whether it was true or false.

The most important finding results from a comparison with Exp. 1; subjects in Exp. 2 were less accurate and less certain of what they were doing. This was apparent at two different levels. First, a greater proportion of subjects in Exp. 2 (42/144, compared with 5/96 in Exp. 1) were inconsistent in their selection of cards to examine. Spontaneous comments from these inconsistent subjects speak chiefly of confusion:

"Right, avoid the consonant, or is it the vowel, I am rather confused" (abstract task).

"I have trouble with this idea, it doesn't seem to make obvious sense, it seems to get more confusing the more I think about it" (contextual task).

However, it is not only the inconsistent subjects who were uncertain –

"I must keep referring back to the rule, I can't seem to remember it"

– notwithstanding that that subject was entirely correct, selecting $\langle R, p \rangle$ throughout and in conformity with all three criteria of logicality. Second, even of those subjects who selected cards with complete consistency, many more failed to make consistently correct evaluations of the truth of the rule (54/86 in Exp. 2, compared with 4/78 in Exp. 1). This subject, who selected all four cards consistently, said, "I've read this rule about three times and it is still not clear to me what it means," and said it was false every time. Note that Exp. 2 employed exactly the same stimuli, instructions, design, and procedure as Exp. 1; even the rules to be tested had the same meaning – only the manner in which they were expressed was different. We conclude that the presence of a negative in a conditional rule makes people less certain what it means and, even relative to their possible misunderstanding, whether it is true or false.

It is logically possible that the two populations of subjects were sufficiently dissimilar for the very salient difference between the two experiments to be due to a difference between subjects, rather than to the use of negatives in the expression of the rule. However, we do not give any credence to this suggestion. The subjects from Weston-super-Mare in Exp. 1 were mostly professional people. The subjects in Exp. 2 all came from a very middle-class area of South Cambridge and, while some were retired, they had previously been chiefly in professional occupations, including university teaching. It needs to be emphasised that no one has previously looked at the consistency of card selection, or of the evaluation of the truth of the rule, when that rule is expressed with negatives, so that there are no prior data with which comparison might be made. While this comparison does need to be replicated, we remark that the much greater difference between the subjects from Weston-super-Mare and the German subjects from Hannover used by Gebauer and Laming (1997) in an equivalent design led to no significant difference at all.

Finally, when performances under the three different rules were aggregated together to give a sample size of 72, there was a highly significant correlation between the two tasks. This correlation was due in part to 10 subjects who failed to find a consistent response combination on either task, and this is understandable – these subjects were simply unsure what to do. However, there was also a significant relationship amongst those subjects who responded consistently on both tasks, with 17 out of 29 selecting equivalent combinations of cards. We had such an idiosyncratic contribution to the understanding of the rules in mind when designing these experiments, and it is a problem for us why this correlation should show strongly in Exp. 2, but not at all in Exp. 1.

Subjects in Exp. 2 were less certain about what they were doing. Many of their spontaneous comments reflected that uncertainty. In addition, several of them said that the contextual task was easier.

"I find the first task I had much easier than this one" (during the abstract task).

"I must say that this rule is much easier than the first one, numbers and letters are a bit confusing" (on the contextual task).

"This is much easier than the numbers and letters problem, there is a little more meaning in this."
None of the subjects in Exp. 1 made such comments, and it needs careful consideration to see what this difference means.

If a task is easy, responses are typically correct and are determined by the task requirements. There is no opportunity for idiosyncratic interpretations to show through, and there is no basis for saying that one easy task is easier than another – the difference is imperceptible. That is the way it was with Exp. 1 (notwithstanding that most subjects misunderstood the rule). However, if a task is difficult, the subject knows it is difficult because of uncertainty what to do. Responses are typically incorrect and therefore poorly determined by the task requirements. They are determined instead by idiosyncratic factors – there is nothing else. Moreover, a difference in subjective uncertainty is appreciable to the subjects; it becomes possible to say of two difficult tasks that one is easier than the other. Thus, the emergence of an idiosyncratic contribution to understanding in Exp. 2 is a plausible consequence of the greater subjective difficulty of negative conditional rules. This needs to be borne in mind in the conduct of future studies.

Discussion

We summarise our findings as follows:

Consistency of card selection

1. As Gebauer and Laming (1997) found, a large proportion of subjects make an entirely consistent selection of cards, consistent with respect to some (mis)interpretation of the rule, over the course of six successive problems. This is true of a contextual task, as well as of the original abstract task.
2. That proportion is somewhat reduced when the rule, the same rule, is expressed with one or two negatives. Even more to the point, notwithstanding that they select cards consistently, many subjects are then uncertain whether the rule as they understand it is true or false.
3. Different selections were made on the abstract and on the contextual task, with more correct contextual selections. This replicates the findings of previous studies, although our contextual proportion correct was lower than that of, say, Griggs and Cox (1982).

We interpret these findings to mean that when subjects select the wrong cards, most often this is because they misunderstand the rule that is to be tested and select, instead, those cards which are *logically* appropriate to testing the rule as they understand it to be. This idea faces, of course, the same criticism which has been levelled at the misinterpretation of syllogisms. Unless there is some independent check on the premises in the syllogism task and of the interpretation of those premises by the subject, the idea of misinterpretation does no more than describe what is observed (Gilhooly, 1982). Such a check is implicit in our experimental paradigm, because performance is observed over six successive problems. There are 15 different card selections that might be made. Suppose our subjects were simply responding at random: the probability of a subject selecting the same combination, any 1 of the 15, on all 6 problems would be 15^{-5} (about 1.3×10^{-6}, if, as a matter of convenience, we ignore the fact that some of those combinations are more frequent than others). It is not the case that the consistent selections we have observed arise by chance.

Different sets of cards are selected, we suggest, because each subject approaches the task with some prior basis in past experience, a different basis for each subject, how rules such as the one in Wason's Selection Task are to be understood. In consequence, the proportion of correct selections – indeed, the entire spectrum of combinations of cards selected – depends on the way in which the rule, the same rule, is expressed. This explains why contextual tasks typically afford a higher proportion of correct selections (Cheng & Holyoak, 1985; Cosmides, 1989; Girotto, Mazzocco, & Tasso, 1997; Kirby, 1994; Sperber, Cara, & Girotto, 1995). Fairley, Manktelow, and Over (1999) have recently underlined this point by presenting the same rule with two different scenarios. One scenario presented the rule as a sufficient condition (in the language of our traffic-light task, "If the traffic light is red, then the car must stop behind the stop line"), while the other scenario made the rule appear a necessary condition ("Only if the traffic light is red, may the car stop behind the stop line"). As a sufficient condition, the rule typically elicited selection of "*R*" and "*p*," while as a necessary condition it chiefly elicited "*G*" and "*s*." We therefore envisage context as constraining the ways in which subjects might understand the rule. Context also applies to the abstract task since, when the original rule (Exp. 1 above) is formulated differently, though still without negatives (Gebauer & Laming, 1997, Exp. 2), the proportion of correct selections is much increased.

The different prior experience of different subjects also explains why they do not all make the same misunderstanding of the rule and do not all select the same wrong combination of cards (their different prior experiences induce them to choose different combinations of cards) and why their performances on the two tasks in Exp. 2 were correlated (the influence of prior experience on the two tasks is similar).

"Logicality"

The interpretation we have placed on the consistency of our subjects' card selections raises the question of what is meant by "logical." We have investigated three criteria.

4. Subjects might select cards consistently, but not always evaluate the rule correctly. Indeed, some subjects do not; but we have not found any systematic correlation of incorrect evaluations between the two tasks.

5. Subjects might turn over additional cards and then realise that they did not need to examine them; they indicated this in our procedure by turning those cards back again. A few subjects did indeed do so but, again, there was no systematic correlation between turning cards back again in the two tasks.

6. Subjects might go on to examine all the potentially relevant cards, even though it could be known from an earlier card that the rule did not hold.

Such behaviour (6) was exceedingly common and might seem to represent no more than a consistent *selection* of cards, as Stenning and van Lambalgen (1999) have suggested; except that Gebauer & Laming (1997) obtained the exact opposite result – they did not observe a single instance of a subject continuing to look after the rule could be known to be false (looking at the "*o*" after discovering a mismatching "*V*"). The difference would seem to lie in the experimental instructions, not the reasoning. Gebauer and Laming (1997, p. 288) told their subjects:

"Bitte drehe aber nur so viele Karten um, wie Du wirklich umdrehen mußt, um entscheiden zu können, ob die Regel richtig oder falsch ist."

("Please turn over only as many cards as you really must in order to be able to decide if the rule is true or false."),

with particular emphasis on the "wirklich umdrehen mußt" (really must turn over). The corresponding instruction here was "Turn over as many cards as you need to discover whether this rule holds; and let me know as soon as you have discovered whether it holds or not. Do not turn over any more cards than you need. If you turn over a card and later decide you didn't need to do so, turn it back again." This, moreover, was merely given to the subjects to read. There was also a natural difference between the experimenters.

The simplest reading is that we are looking at two different ways of going about the task. Gebauer and Laming's (1997) subjects evaluated the rule after each individual card was examined, which is what was intended. Our subjects examined a selection of cards, appropriate to their understanding of the rule, and only then decided whether the rule held or not. Some subjects did indeed turn some cards back again as not needing to be examined, but never the "*o*" following a mismatching "*V*." If subjects were merely *selecting* cards consistently, then they would have had no basis on which to decide whether the rule was true or false, but in Exp. 1 there were only four false evaluations of the rule amongst 78 subject-performances with consistent selections throughout (though Exp. 2 is a different story). With a view to future work, we recommend that the requirement to terminate the problem as soon as the rule can be known to be false (as in Gebauer & Laming, 1997) be retained. In effect, this requires the subject to make a decision whether the rule holds or not after turning each individual card, rather than a single decision when all potentially relevant cards have been examined. A succession of decisions based on single cards will be more informative.

Asking subjects to turn cards over and to discover whether the rule holds or not is the critical innovation in procedure. It greatly enhances the yield of information; it makes it feasible to present a series of problems, and to distinguish between misunderstanding of the rule and inaccuracies in reasoning from that (mis)understanding. The undersides of the cards provide the subject with "feedback," but we emphasise that this is feedback only with respect to whether the rule (as understood by the subject) holds or not and says nothing about whether that understanding is correct. Allowing subjects thereafter to turn those cards back again which did not need to be examined might appear controversial. However, we emphasise, again, that the turning back does not provide the subject with any additional feedback – the feedback is provided when the cards are turned over the first time. What the turning back does do is to provide the experimenter with additional insight into the subject's mental processes.

Failure to evaluate the rule correctly (4) was related to the presence and number of negatives. Why do negatives lead to confusion? Our experiment was not designed to speak to that question, and any suggestion we make here is no more than guidance for some further investigation. However, that further investigation might usefully ask whether the rule is initially understood as an affirmative statement (i.e., If *V*, then *e*) and then inverted, a separate inversion for each negative. Such a procedure was suggested long ago by Wason (1961). The inversion sometimes fails – this would explain why two negatives are more difficult than one and why there are almost no failures of evaluation in Exp. 1.

We now turn to other possible interpretations of these results.

Matching bias

Following the results of some experiments with truth tables, Evans (1972) proposed that subjects failed on Wason's selection task as well as on truth tables because they tended to select those cards named in the statement of the rule (i.e., *V*, *e*/*R*, *s*). This hypothesis was labelled *matching bias*. Matching bias is thought (Evans, 1989) to be based on two heuristics: an "If" heuristic which causes subjects to choose the card (*V*) corresponding to the antecedent in the rule and a "Not" heuristic which causes subjects to ignore negations and examine, instead, the cards which are explicitly mentioned as negated (including both *V* and *e* in an affirmative rule). On this basis, *e* is addressed by the "Not" heuristic only, while *V* is addressed by both and is therefore selected

more frequently than *e*. It means that matching bias shows up chiefly in the selection of the "true consequent" (*e*; Evans, 1989, p. 57).

The obvious test of matching bias as a general description of performance on Wason's task is to make a comparison with rules expressed with negatives. Evans and Lynch (1973) made that comparison for the abstract task; Griggs and Cox (1983) repeated their work with thematic conditionals. Griggs and Cox claimed that "logical processing" accounted for more of the data than any other strategy, especially in meaningful contexts, and on that basis suggested that matching bias was a last resort when responding to less meaningful material. More recently, Evans, Clibbens, and Rood (1996) have concluded that "matching bias" applies only to implicit negation (that is, to "*o*" in the rule "If *V*, then *e*"). When the rule is expressed with explicit negations ("If *V*, then $\neg o$"), matching bias largely disappears.

Our Exp. 2 showed no significant difference between the combinations of cards selected depending on the location of the negative(s) in the rule (and therefore which elements were explicitly negated) and at best a slight difference compared with Exp. 1. The "Not" heuristic is therefore contraindicated. That is to say, while "matching bias" describes the most frequent selections in the abstract task, it does not extrapolate to negated rules (as Evans, Clibbens, & Rood, 1996, have already reported). It does not extrapolate, either, to our contextual task, where the most common selection is ⟨all 4 cards⟩, and the second most common, $\langle R, p\rangle$, which is correct. In short, "matching bias" does no more than describe the most common selection $\langle V, e\rangle$ in the abstract task and, as an adjunct, the second most common $\langle V\rangle$ as well.

However, these results do admit a simple explanation. Gebauer and Laming (1997, Table 4) noted two component misinterpretations of the rule, reading "top/underneath" for "one side/other side" and bi-conditional for simple conditional. If these two component misinterpretations are combined, the logical selection of cards is $\langle V, e\rangle$, which is the principal pattern addressed by "matching bias." Table 6a sets out the frequencies of the most common card selections listed in Table 2, categorising them according to these two component misreadings, while Table 6b presents the frequencies of shifts in these components by those subjects who made just one change of selection during the six problems.

Looking at the top four rows only, there is only a small difference between the three sets of abstract data ($G^2 = 13.963$ with 6 *df*, $p < 0.05$) and none between the two sets of contextual data ($G^2 = 1.101$ with 3 *df*); that is to say, reformulating the rule with negatives makes little difference. (To be more precise, visual inspection of Table 6a suggests that the abstract data from Exp. 2 is like the contextual data from either experiment and different from other abstract data. This is confirmed by analysis. Comparing now all five rows in Table 6a: cols 1 & 2 [abstract data from Gebauer and Laming, 1997, & our Exp. 1], $G^2 = 4.212$ with 4 *df*; cols 3 to 5 [contextual task in Exp. 1 and both tasks in Exp. 2], $G^2 = 6.465$ with 8 *df*; residual between these two groups, $G^2 = 42.794$ with 4 *df*). Continuing to look at the top four rows only, the two component misinterpretations occur independently in each of the two tasks (abstract task:

Table 6a Component errors of interpretation

Component errors	Card selection	Gebauer & Laming, 1997	Experiment 1		Experiment 2		Totals
			Abstract	Contextual	Abstract	Contextual	
Top/underneath	*V*/*R*	4	7	6	6	5	28
Top/underneath & bi-conditional	*V*, *e*/*R*, *s*	11	17	5	7	4	44
Correct	*V*, *o*/*R*, *p*	1	1	9	4	13	28
Bi-conditional	All 4	2	6	13	14	11	46
Other		0	4	10	11	12	37
Total		18	35	43	42	45	183

Table 6b Switches in interpretation

Component errors	Gebauer & Laming, 1997	Experiment 1		Experiment 2	
		Abstract	Contextual	Abstract	Contextual
Bi-conditional to conditional	3 *V*, *e* → *V*				
Top/underneath to 1 side/other side	1 *V*, *e* → All 4	7 *V*, *e* → All 4	1 *R*, *s* → All 4	1 *V* → *V*, *o*	2 *R* → *R*, *p*
Conditional to bi-conditional & Top/underneath to 1 side/other side			1 *R* → All 4		1 *R* → All 4
Other		3	1		

$X^2 = 1.127$; contextual task: $X^2 = 0.287$; both with 1 *df*), but the incidences of each component do differ between the tasks (''Top/underneath'' for ''one side/other side'': $X^2 = 17.418$ with 1 *df*, $p < 0.001$; bi-conditional for simple conditional: $X^2 = 6.907$; with 1 *df*, $p < 0.01$). The incidence of the bi-conditional misinterpretation is reduced in our contextual task and ''top/underneath'' is *much* reduced relative to the abstract task. The consequence is a greatly reduced incidence of $\langle R, s\rangle$ selections relative to $\langle V, e\rangle$. We suggest that this reduction in the frequencies of these component misinterpretations is due to constraints implicit in the context.

Card selections on the first problem

Our Exp. 2 failed to show any differences in the selection of cards between the three rules with negatives, but that analysis focused on the pattern of selection over all six problems. To facilitate comparison with previously published results, Table 7 sets out the frequencies for each individual card on the first problem only. Comparison of these frequencies substantially confirms the previous analysis. Overall, there is a highly significant difference ($G^2 = 77.553$ with 28 *df*, $p < 0.001$). Decomposition of that G^2 statistic reveals a difference between the two experiments ($G^2 = 18.885$ with 4 *df*, $p < 0.001$) and a difference between the abstract and contextual tasks in Exp. 1 ($G^2 = 22.856$ with 4 *df*, $p < 0.001$). In Exp. 2 there is a slight difference between the frequencies of selection between the three rules in the contextual task ($G^2 = 16.164$ with 8 *df*, $p < 0.05$), but no corresponding difference for the abstract task ($G^2 = 11.929$ with 8 *df*) and no difference between the two tasks overall ($G^2 = 7.718$ with 4 *df*). The results of this analysis differ only slightly from what was reported above, so the question arises: Why do we fail to find any difference between the three rules with negatives, when previous investigators have systematically reported confirmatory findings (see Oaksford & Chater, 1994, Table 3)?

All of the studies included in Oaksford and Chater's (1994) meta-analysis (Evans & Lynch, 1973; Manktelow & Evans, 1979; Oaksford & Stenning, 1992) used within-subject designs, each subject attempting all four rules. Those four rules have to be administered in some order and, notwithstanding that that order was randomised, for three of the problems (three-quarters of the data) the subjects have immediately prior experience of a related, but slightly different rule – the context is different. The argument has already been put that such changes of context make a difference, so the rigorous comparison has to be between the frequencies in Table 6 and the corresponding frequencies from *the first problem only* attempted by the subjects in these three experiments – data we do not have.

To put this discussion of matching bias into perspective, compare Gebauer and Laming's (1997) reformulation of the rule:

''If there is a vowel on top of the card, there is an even number underneath, and if there is an odd number on top of the card, there is a consonant underneath.''

This reformulation mentions all four cards (the ''matching bias'' prediction), but selection of ⟨all 4 cards⟩ was not observed. Instead, three of the five subjects with the rule expressed in this way made correct selections on all six problems presented, and that proportion (3/5) was sufficient to establish an improvement significant at 0.001 with respect to the original form of the rule used in Gebauer and Laming's Exp. 1.

Mental models

Mental models theory provides a concise notation in which to represent the internal mental processes hypothesised in the mind of the problem-solver. Johnson-Laird and Byrne (1991) and Johnson-Laird (1995) have proposed a model for Wason's selection task within the mental models framework. There are three stages to the reasoning process:

1. First, the rule is represented by a mental model; this might be accomplished in more than one way. ''...When content and context are neutral, sometimes a conditional is taken to imply its converse [bi-conditional], and sometimes not: people are neither consistent with one another nor from one occasion to another.'' (Johnson-Laird & Byrne, 1991, p. 46).

2. The mental representation is ''fleshed out'' to yield a set of possible models of the rule and an initial conclusion.

3. The subject searches for counterexamples to invalidate the different models of the rule. ''Errors occur, according to the theory, because people fail to consider all possible models of the premises. They therefore fail to find counterexamples to the conclusion that they derive

Table 7 Frequencies of selection of individual cards on first problems only

Rule	Experiment 1		Experiment 2					
	If V, then e	If R, then s	If V, then $\neg o$	If R, then $\neg p$	If $\neg C$, then e	If $\neg G$, then s	If $\neg C$, then $\neg o$	If $\neg G$, then $\neg p$
V/R (True Antecedent)	45	40	22	24	19	21	19	20
C/G (False Antecedent)	35	23	8	9	16	9	16	11
e/s (True Consequent)	12	17	13	4	13	12	12	12
o/p (False Consequent)	12	29	13	15	14	13	18	17

from their initial models'' (Johnson-Laird & Byrne, 1991, p. 39).

We remark, straightaway, that this is a much more elaborate scenario than can be supported simply by the data from Wason's selection task. It is undetermined in a sufficient number of places to be able to accommodate almost everything and predict nothing. There are, however, four critical points.

1. Our principal experimental finding is that a majority of our subjects (68%) selected cards for examination in complete logical conformity with some misunderstanding of the rule to be tested; the inconsistencies of the remainder, on the other hand, cannot be reliably assigned to logical error rather than to plain confusion over the meaning of the rule. Our experimental data therefore exemplify the thesis advanced by Henle (1962): ''I have never found errors which could unambiguously be attributed to faulty reasoning'' (Henle, 1978). Johnson-Laird (1983, pp. 25–26) roundly rejected that idea.

2. The most common misreading of Wason's rule is to substitute a bi-conditional for the simple conditional. As Table 6a shows, those of our subjects who misread the rule as a bi-conditional did so consistently, contrary to the dictum from Johnson-Laird and Byrne (1991, p. 46) above.

3. However, nearly as common was to substitute ''top/underneath'' for ''one side/other side.'' Table 6b shows that most of the changes of mind amongst those subjects who changed their card selection just once during the six problems concerned these two component misinterpretations. While the mental models theory has the capability to represent a biconditional interpretation of the rule (this particular misinterpretation has long been remarked), there is, as yet, no facility for distinguishing between the two sides of the cards. Yet this substitution of ''top/underneath'' for ''one side/other side'' answers the problem of performance on Wason's task as acutely stated by Johnson-Laird (1983, p. 30): ''The serious mistake arises with the card bearing the odd number. Very few subjects elect to turn it over, and this sin of omission is puzzling because if the card has a vowel on its other side then the generalization is blatantly false.'' If we might express this theoretical error in terms of the mental models theory itself, Johnson-Laird (1995) failed to ''flesh out'' his theory of the selection task to distinguish between the top and the underside of the cards. As Johnson-Laird and Byrne (1991, p. 39) remark, ''Errors occur, according to the theory, because people fail to consider all possible models of the premises.''

4. Our investigation of ''logicality'' has uncovered other facets of behaviour on the selection task (incorrect evaluation of the rule; turning cards back again as not needing to be examined; examining all potentially relevant cards, notwithstanding that the rule can already be known not to hold) which also lack any representation in the mental models theory. It is not that the authors have failed to incorporate appropriate assumptions into their model, but, rather, that the mental models notation has not, as yet, been developed to represent the underpinnings of these particular facets of performance.

A ''fleshed out'' mental model is, of course, an understanding, or a misunderstanding, of the rule in the sense in which we use the term. However, our conception of ''misunderstanding'' subsumes a wider category of potential interpretations than the mental models theory, as it is at present formulated, can accommodate (though the mental models theory could itself be ''fleshed out'' to the extent required). More critical, however, is this difference of conception: We envisage that the misunderstanding of the rule is immediate, preconscious. While it may be that subjects perceive the rule exactly as stated and arrive at their apparent misunderstandings by a process of conscious, but incorrect, reasoning, we see each subject's misunderstanding of the rule as primary and reasoning beginning from that datum.

Correlations with related tasks

Green (1995b) has suggested that one might explore the process of card selection in more detail by studying variants of the selection task – card specification, thinking aloud, and externalisation – and, in particular (Green, 1995a), that one might discover how a subject really understands the rule by asking for the identification of counter-examples. For example, Green and Larking (1995) set their subjects three tasks in order: (a) to envisage a counter-example, (b) to determine which cards could provide it, and (c) the selection task itself. The accumulated results of these and similar experiments (Green, 1995a, b) are, for the most part, disappointing. The results do not open the card selection process up to understanding, and we will suggest why this might be so.

Green and his colleagues are, of course, looking for elaborations of the experimental procedure which will increase the proportion of correct $\langle V, o \rangle$ selections. Let us put a slightly different gloss on that enterprise and say that they are looking for correlations between the proportions of correct performances on the selection task and on its close relatives. Our results in this paper and in Gebauer and Laming (1997) show that such correlations are strongest between the selection task and itself (of course), and they focus not so much on the proportion of correct $\langle V, o \rangle$ selections, but around the *entire pattern* of card selections. Suppose the relationships between the selection task and its relatives – card specification, thinking aloud, and externalisation – to be re-examined from the standpoint that each subject has an idiosyncratic understanding of what is to be done, a different understanding by different subjects. We do not have the data and cannot say that anything will shake out, but we think it worth trying.

Conditional inference task

Our version of the selection task brings out a close relationship with the conditional inference task, sometimes known as "syllogisms." Evans and Handley (1999) have studied conditional inference in a Wason-selection-task format to address this question: "Is the representation of conditionals task-dependent?" This question arises from what they call the "double negation effect," which is "a tendency to endorse more often inferences for which conclusions are negative than those for which conclusions are affirmative in form. However, it has recently been shown (Evans et al., 1995) that this bias – although very reliable – is consistently manifest only on the two forms of inference involving denial: MT [modus tollens] and DA [Denial of the Antecedent]" (Evans & Handley, 1999, p. 742). They go on to assert that the mental models theory "can account for the double negation effect in conditional inference and matching bias on the selection task, but only by proposing a difference in how the conditional is represented on each task" (Evans & Handley, 1999, p. 746). Without going into their argument, we shall point out a much simpler resolution of the problem.

There are two sources of error in a reasoning task. Responses might be incorrect because of (a) a misperception of what the task consists of, or (b) a failure to reason accurately from that (mis)perception. Wason's selection task in the customary version used by Evans and Handley (1999) is an exception. If subjects are asked merely to indicate which cards would need to be examined, their card selections indicate only their understanding of the rule to be tested and say nothing about their accuracy of reasoning from that understanding. If, however, subjects are asked to turn the cards over and announce whether the rule holds or not, they are making conditional inferences. That is to say, the conventional selection task requires only Stage 1 of a reasoning model (interpretation of the rule), while conditional inference requires Stage 2 (reasoning from that rule). The theoretical conflict which Evans and Handley (1999) address is illusory.

Failure to evaluate the rule

Our version of the selection task is also related to the truth-table task, introduced by Johnson-Laird and Wason (1970b). In our task there are four cards (*V*, *C*, *e*, *o* / *R*, *G*, *s*, *p*). The subject selects certain of these cards and examines the underside, leading to these possible pairings:

V, *e*/*e*, *V*	*R*, *s*/*s*, *R*	Antecedent and consequent are both "true."
V, *o*/*e*, *C*	*R*, *p*/*s*, *G*	Antecedent is "true," consequent is "false."
C, *e*/*o*, *V*	*G*, *s*/*p*, *R*	Antecedent is "false," consequent is "true."
C, *o*/*o*, *C*	*G*, *p*/*p*, *G*	Antecedent and consequent are both "false."

Here, "true" means "mentioned in the statement of the rule." The subject then implicitly, for some cards explicitly, states whether or not that particular pair conforms to the rule. This is truth-table evaluation for those cards which the subject selects. Those cards which are not selected are, by default, classified as irrelevant.

Truth tables with negative conditionals have been cited (Evans 1972; Evans 1998; Evans, Clibbens, & Rood, 1996; Evans & Lynch, 1973) as support for the idea of matching bias. Our finding in Exp. 2 is that subjects with negative conditionals frequently get the evaluation wrong. It is especially noteworthy that in Exp. 2 many subjects, otherwise selecting cards with complete consistency, were uncertain whether the rule, even as they understood it, was true or false. This might look, at first sight, to be direct empirical verification of Evans' "Not" heuristic – but it is not. The idea, neglect of the negative in the rule, is the same, but the point of application is different – selection of a card in the "matching bias" story and evaluation of the truth of a rule in our experiment. What our result most immediately means is that negatives in a conditional rule simply confuse people and should not be used in any practical application that matters.

Conclusion

Subjects fail Wason's selection task principally because they misunderstand the rule they are asked to test. This misunderstanding has a profound implication for the study of reasoning. In any reasoning task, there are two sources of error. Responses might be incorrect because of (a) a misperception of what the task consists of, or (b) a failure to reason accurately from that (mis)perception. A model for the task must therefore incorporate *two* stages to accommodate these distinct sources of error, and the evaluation of such a model requires the experiment to identify how each individual subject perceives the task (in the selection task, how each individual subject interprets the rule). When subjects in Wason's selection task are asked merely to indicate which cards need to be examined, their card selections indicate only their understanding of the rule to be tested and say nothing about their accuracy of reasoning from that rule. That is to say, Wason's task, as it has conventionally been implemented, *involves no reasoning* (except to the extent that "reasoning" can be subsumed in "understanding"). It is not surprising that models which do no more than seek some specific defect in reasoning cannot accommodate performance on Wason's task.

Acknowledgements The experiments were undertaken by the first author as part of her work for an M.Phil. at the University of Cambridge, 1997–98, under the supervision of the second author. We thank Keith Stenning and Walter Schaeken for their comments on an earlier draft of this paper.

References

Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: Wiley.

Bracewell, R. J., & Hidi, S. E. (1974). The solution of an inferential problem as a function of stimulus materials. *Quarterly Journal of Experimental Psychology*, *26*, 480–488.

Bradley, D. R., Bradley, T. D., McGrath, S. G., & Cutcomb, S. D. (1979). Type I error rate of the chi-square test of independence in R × C tables that have small expected frequencies. *Psychological Bulletin*, *86*, 1290–1297.

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

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.

Evans, J. St. B. T. (1972). Interpretation and matching bias in a reasoning task. *Quarterly Journal of Experimental Psychology*, *24*, 193–199.

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

Evans, J. St. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking and Reasoning*, *4*, 45–82.

Evans, J. St. B. T., Clibbens, J., & Rood, B. (1995). Bias in conditional inference: Implications for mental models and mental logic. *Quarterly Journal of Experimental Psychology*, *48A*, 644–670.

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

Evans, J. St. B. T., & Handley, S. J. (1999). The role of negation in conditional inference. *Quarterly Journal of Experimental Psychology*, *52A*, 739–769.

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

Fairley, N., Manktelow, K., & Over, D. (1999). Necessity, sufficiency, and perspective effects in causal conditional reasoning. *Quarterly Journal of Experimental Psychology*, *52A*, 771–790.

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

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

Gilhooly, K. J. (1982). *Thinking directed, undirected, and creative*. London: Academic press.

Girotto, V., Mazzocco, A., & Tasso, A. (1997). The effect of premise order in conditional reasoning: A test of the mental model theory. *Cognition*, *63*, 1–28.

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

Green, D. W. (1995b). The abstract selection task: Thesis, antithesis, and synthesis. In S. E. Newstead and J. St B. T. Evans (Eds.), *Perspectives on thinking and reasoning*. (pp. 173–188) Hove UK: Erlbaum.

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

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

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

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

Henle, M. (1978). Foreward to R. Revlin & R.E. Mayer (Eds.), *Human reasoning*. Washington, DC: Winston.

Ishihara, S. (1967). *The series of plates designed as a test for colour-deficiency*. Tokyo: Kanehara Shuppan; London: Lewis.

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

Johnson-Laird, P. N. (1995). Inference and mental models. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning*. (pp. 115–146) Hove, UK: Erlbaum.

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

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

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

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

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association*, *73*, 253–263.

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

Norusis, M. J. (1985). *SPSSx advanced statistics guide*. New York: McGraw-Hill.

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

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

Revlin, R., & Leirer, van O. (1980). Understanding quantified categorical expressions. *Memory and Cognition*, *8*, 447–458.

Smalley, N. S. (1974). Evaluating a rule against possible instances. *British Journal of Psychology*, *65*, 2, 293–304.

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

Stenning, K., & van Lambalgen, M. (1999). Is psychology hard or impossible? Reflections on the conditional. In J. Gerbrandy, M. Marx, M. de Rijke, & Y. Venema (Eds.), *Liber amicorum for Johan van Bentham's 50th birthday* (pp. 1–29). Amsterdam: Amsterdam University Press.

Taplin, J. E. (1971). Reasoning with conditional sentences. *Journal of Verbal Learning and Verbal Behaviour*, *10*, 219–225.

Taplin, J. E., & Staudenmayer, H. (1973). Interpretation of abstract conditional sentences in deductive reasoning. *Journal of Verbal Learning and Verbal Behaviour*, *12*, 530–542.

Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology*, *52*, 133–142.

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology*, (p. 135–151). Harmondsworth, UK: Penguin.

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., & Evans, J. St. B. T. (1975). Dual processes in reasoning? *Cognition*, *3*, 141–154.

Wason, P. C., & Johnson-Laird, P. N. (1970). A conflict between selecting and evaluating information in an inferential task. *British Journal of Psychology*, *61*, 509–515.