REASONING ABOUT A RULE

BY

P. C. WASON

From Psycholinguistics Research Unit, University College London

Two experiments were carried out to investigate the difficulty of making the contrapositive inference from conditional sentences of the form, "if P then Q." This inference, that not-P follows from not-Q, requires the transformation of the information presented in the conditional sentence. It is suggested that the difficulty is due to a mental set for expecting a relation of truth, correspondence, or match to hold between sentences and states of affairs. The elicitation of the inference was not facilitated by attempting to induce two kinds of therapy designed to break this set. It is argued that the subjects did not give evidence of having acquired the characteristics of Piaget's "formal operational thought."

INTRODUCTION

This investigation is concerned with the difficulty of making a particular type of inference from conditional sentences, statements of material implication of the form, "if P then Q" ($P \supset Q$).

Within the propositional calculus the truth table for the conditional counts the following combinations of the components of the sentences as true: PQ, $\bar{P}Q$, $\bar{P}\bar{Q}$ (where $\bar{P}$ stands for not-P and $\bar{Q}$ for not-Q), and only the one combination, $P\bar{Q}$, as false. It follows that only values of P and values of $\bar{Q}$ allow a valid inference. In other words a valid inference depends crucially upon the *possibility* of meeting the falsifying contingency, $P\bar{Q}$. It rests simply upon the denial of this contingency. $\bar{P}$ comes out true whether it is associated with Q or $\bar{Q}$,and Q comes out true whether it is associated with P or $\bar{P}$. For example, suppose someone says, "if John is a stockbroker, then he is bound to read the *Financial Times*." If John isn't a stockbroker, or if he does read the *Financial Times*, the conditional statement (to the logician) is inevitably true. But if he is a stockbroker, or if he never reads the *Financial Times*, then there are grounds for verifying or falsifying the conditional statement. Hence, logically, only two forms of inference are valid: "$P \supset Q$ and P . . . Q" (*modus ponens*) and "$P \supset Q$ and $\bar{Q}$ . . . $\bar{P}$" (*modus tollens* or contrapositive). The other two forms of inference, "$P \supset Q$, and $\bar{P}$ . . . $\bar{Q}$" (denial of the antecedent) and "$P \supset Q$ and Q . . . P" (affirmation of the consequent) are fallacious.

In a task consisting of sentences expressed in everday terms the author (Wason, 1964) has shown that the affirmation of the consequent occurs significantly more frequently than the denial of the antecedent. In an abstract task the occurence of all four types of inference has been investigated in a pilot study (Wason, 1966). The subjects were presented with the following sentence, "if there is a vowel on one side of the card, then there is an even number on the other side," together with four cards each of which had a letter on one side and a number on the other side. On the front of the first card appeared a vowel (P), on the front of the second a consonant ($\bar{P}$), on the front of the third an even number (Q), and on the front of the fourth an odd number ($\bar{Q}$). The task was to select all those cards, but only those cards, which would have to be turned over in order to discover whether the experimenter was lying in making the conditional sentence. The results of this study, and that of a replication by Hughes (1966), showed the same relative frequencies of cards selected. Nearly all subjects select P, from 60 to 75 per cent. select Q, only a minority select $\bar{Q}$ and hardly any select $\bar{P}$. Thus two errors are committed: the consequent is

fallaciously affirmed and the contrapositive is withheld. This type of task will be called henceforth the "selection task." These errors seem to be enduring and deep. Hughes has shown that they often persist for 15 trials, even when the subjects turn over all the cards after each trial and evaluate the conditional sentence with respect to them.

A theory to explain these results has been postulated (Wason, 1966). It rests on two assumptions. The first assumption is that individuals are not constrained by the rules of the propositional calculus. They implicitly assume that a conditional sentence can have three outcomes or truth values. PQ is true, P$\bar{Q}$ is false and $\bar{P}$ with either Q or $\bar{Q}$ is irrelevant. This interpretation is not original: it has been debated in the history of logic since the Stoic and Megarian schools. The assumption explains why the consequent is affirmed—Q is selected in order to see whether it is associated with P making the conditional true. It also explains why the antecedent is so infrequently denied—$\bar{P}$ is irrelevant to the truth or falsity of the sentence.

The second assumption explains the infrequency of the contrapositive inference—why $\bar{P}$ is so seldom deduced from $\bar{Q}$. This assumption is that individuals are biased, through a long learning process, to expect a relation of truth, correspondence or match to hold between sentences and states of affairs. In adult experience truth is encountered more frequently than falsity, and we seldom use a proposition or judgement that something is false in order to make a deduction. The semantic concept of falsity is logically equivalent to the syntactic concept of negation, and it has been shown that both cause difficulty when sentences have to be evaluated or constructed (Wason, 1959, 1961; Wason and Jones, 1963). Both concepts are relevant to the selection task. A value of $\bar{Q}$ represents a mismatch between a state of affairs visible on the card and a clause in the conditional sentence. This mismatch has to be recognized as such by transforming the information in the relevant clause from Q to $\bar{Q}$. In other words, a state of affairs (x) has to be seen not simply as x but as $\bar{Q}$. In doing this the individual presumably makes a judgement of falsity by uttering a covert negative sentence to himself. The difficulties involved in doing all this are assumed to be sufficient to account for the relative failure to select $\bar{Q}$.

According to this theory the affirmation of the consequent—the deduction of P from Q—is a plausible inference. But the withholding of the contrapositive remains irrational and is consequently a factor of much greater interest. The present investigation is concerned with the effects of therapies designed to correct the bias towards truth or correspondence, and thus facilitate the elicitation of the contrapositive inference.

EXPERIMENT I

The projection of falsity

It was predicted that if individuals were to "project falsity," i.e. to say what values, if any, associated with the given values, P, $\bar{P}$, Q, $\bar{Q}$, would make a conditional sentence *false*, then they would more readily select $\bar{Q}$ to determine whether the sentence was in fact true or false. P. is the only value, associated with $\bar{Q}$, which could make a conditional sentence false. Hence if individuals were to project P on to a value of $\bar{Q}$, they might be more likely to select $\bar{Q}$ as informative.

Method

Design. In the experimental group the subjects first carried out a selection task. They decided which of four cards (P, $\bar{P}$, Q, $\bar{Q}$) would enable them to determine whether a given conditional sentence was true of false, if they were to know the values on the back of the cards. They were then invited to say what values, if any, on the back of the cards would render the sentence false. They then revised their initial selection of the cards, if

they wished to do so, and finally evaluated the sentence with respect to the values on both sides of all four cards, i.e. they turned over each card and evaluated the sentence as true or false with respect to it. In the control group the procedure was similar, but instead of projecting falsity the subjects were simply asked to think again about their initial selections, i.e. to revise them, "because people often do this task too quickly and get it wrong."

Subjects. Thirty-six first year psychology and statistics students of University College London.

Material. Two conditional sentences were typed on separate cards (5 × 3 in.): (1) "If there is a D on one side of any card, then there is a 3 on its other side," (2) "If there is a 3 on one side of any card, then there is a D on its other side." These two sentences were used as a control for the order in which the two items were mentioned. Associated with each sentence were four cards (2 × 2 in.). The cards associated with sentence (1) had the following letters and numbers on their front, and (in brackets) the following on their back: D(3), 3(K), B(5), 7(D). The following cards were associated with sentence (2): 3(D), K(3), 5(B), D(7). It will be appreciated that both sets of cards conform to only one of the two possible combinations of values of the antecedent and consequent: P($\bar{Q}$), Q($\bar{P}$), $\bar{P}$($\bar{Q}$), $\bar{Q}$($\bar{P}$). The other combination was not used: P($\bar{Q}$), Q(P), $\bar{P}$(Q), $\bar{Q}$($\bar{P}$).

Procedure. The subjects were allocated alternately to the groups and tested individually. Half the subjects in each group were tested with sentence (1) and half with sentence (2). The sentence was placed on the desk and the four test cards were placed in a line in a random order face upwards in front of the sentence. The subjects were told that cards with letters on their front had numbers on their back and *vice versa*. In the selection task the experimenter pointed to each card in turn and asked the subject whether knowing what was on the other side would enable him to find out whether the sentence was true or false. During the projection of falsity, in the experimental group, the experimenter pointed to each card and asked the subject to name a letter (or number) on the other side which would make the sentence false, or to say "none," if none would do so.

RESULTS

Two subjects, both in the control group, seemed unable to comply with the instructions and were hence rejected.

Tables I and II show the frequency of the responses in the different phases of the task for the experimental and control groups respectively. It will be noted immediately that, by comparing the frequencies in the diagonal cells, there is a greater tendency for the initial and revised selections of the cards to conform in the control group than in the experimental group. Inspection shows that the frequency of selecting $\bar{Q}$ increases over the task from five to eight in the experimental group and from two to three in the control group. The prediction that the selection of $\bar{Q}$ would be facilitated significantly in the experimental group is not confirmed.

Table I shows that in the experimental group 13 subjects did not select $\bar{Q}$ initially, and that of these, five did not project P on to $\bar{Q}$ and eight did project it. But of these eight, only three included a value of $\bar{Q}$ in their revised selection. Thus the therapy of falsifying the values cannot always be induced, and even when it is induced it is by no means effective.

Table I shows that the most frequent projection of falsity was on to all four values (P, $\bar{P}$, Q, $\bar{Q}$). But only one of the six subjects responsible for doing this subsequently selected all four values in their revised selection. It is evident that in these cases falsification did not render all the values acceptable for testing the truth or falsity of the sentence. The invitation to project falsity in these cases seemed to result in an arbitrary or indiscriminate response which had no bearing on subsequent behaviour.

It will be noted that P and Q is selected with the greatest frequency in both the experimental group (eight cases) and in the control group (10 cases).

TABLE I

FREQUENCY OF INITIAL SELECTIONS, PROJECTIONS OF FALSITY AND REVISED SELECTIONS IN THE EXPERIMENTAL GROUP

Initial selection	*Projection of falsity*	Revised selection							
		PQ	P	PQ̄	PP̄Q	PP̄QQ̄	PQ̄	N	N
P Q	P Q P P̄ Q Q̄	2 4				1 1		3 5	8
P	P Q P P Q̄ P P̄ Q Q̄	1 1 1					1	1 1 1 1	4
P Q Q̄	P P Q̄ P Q Q̄		1	1 1				1 1 1	3
P P̄ Q	P Q̄						1	1	1
P P̄ Q Q̄	P̄ Q̄ P Q Q̄						1 1	1 1	2
	N	9	1	2	0	2	4	18	18

TABLE II

FREQUENCY OF INITIAL AND REVISED SELECTIONS IN THE CONTROL GROUP

Initial selection	*Revised selection*						
	PQ	P	PQ̄	PP̄	PQQ̄	PP̄QQ̄	N
P Q	9					1	10
P	1	2					3
P Q̄			1				1
P P̄				1			1
P Q Q̄					1		1
P P̄ Q Q̄							0
N	10	2	1	1	1	1	16

Table III is the frequency distribution of the evaluation of the contingencies as true, false or irrelevant for the combined groups.

It will be noted that the author's theory is corroborated with respect to the evaluation of the PQ, PQ̄ and P̄Q̄ contingencies, but is refuted with respect to the evaluation of the P̄Q contingency. This was evaluated as making the sentence false by 22 subjects and as being irrelevant by only 10 subjects. This result is particularly

TABLE III

FREQUENCY OF THE EVALUATIONS IN THE COMBINED GROUPS

PQ	$\bar{P}Q$	$P\bar{Q}$	$\bar{P}\bar{Q}$	N
t	f	f	i	15
t	i	f	i	9
t	f	f	t	3
t	t	f	t	2
t	f	f	f	2
i	f	f	i	1
t	f	i	i	1
t	i	t	i	1
				34

t = true, f = false, i = irrelevant.

interesting because, on the revised selection, $\bar{P}$ was selected only four times out of 34, and yet, when it is associated with Q there is a much greater tendency for it to be judged as relevant to the falsity of the sentence.

This experiment has shown the inadequacy of projecting falsity as a therapy for the elicitation of the contrapositive inference, and that some subjects cannot even perform the therapeutic exercise. Two further pilot studies were carried out which revealed one interesting phenomenon. In the first study the four values were presented on separate trials and a few subjects, when presented with $\bar{P}$ and $\bar{Q}$, said they did not need to turn these over because they already falsified the conditional sentence. In the second study the subjects were asked to pick out only those values which "could break the rule" (i.e. falsify the conditional sentence). Four subjects selected *only* values of $\bar{P}$ and $\bar{Q}$ and refused to turn them over because they claimed this was useless. "It doesn't make any difference—the two I have chosen do break the rule." "There is no rule regarding that card ($\bar{P}$)." Thus in a small minority of subjects the concept of something following a rule appears to be inadequately conceived, for to know what could follow a rule is to know what could break that rule.

EXPERIMENT II

The restricted contingency programme

It was predicted that if subjects were initially allowed to evaluate examples of the four contingencies with respect to a given conditional sentence, and were, in addition, told that only one contingency falsified the sentence, then they would subsequently select values of $\bar{Q}$ within the same task to a greater extent than those who had not had this experience. The term, "restricted," is used because the intention of this programme was not to teach the truth table for the conditional, but to make the subject aware that only the $P\bar{Q}$ contingency falsified the sentence. It was reasoned that if the subject knows in advance that $\bar{Q}$ is crucial for falsification, then he might select it as potentially informative.

Method

Design. The experimental group first received the programme and then carried out a selection task with the same material. The control group carried out a selection task without receiving the programme.

Subjects. Twenty-six first year psychology students at the University of Edinburgh.

Material. Two conditional sentences were typed on separate cards similar to those used in Experiment I. (1) "If there is a square on one side of the card, then there is a red scribble on the other side," (2) "If there is a red scribble on one side of a card, then there is a square on the other side." Four programme cards were prepared for the experimental group which had the following stimuli on either side: (*a*) square, yellow scribble, (*b*) square, red scribble, (*c*) rectangle, red scribble, (*d*) hexagon, brown scribble. Eight similar cards were prepared for the selection task, the items mentioned first being on the front of the cards, and the items mentioned second (in brackets) being on the back: (*a*) square (red scribble), (*b*) square (brown scribble), (*c*) red scribble (square), (*d*) red scribble (hexagon), (*e*) green scribble (rectangle), (*f*) parellelogram (yellow scribble), (*g*) triangle (red scribble), (*h*) blue scribble (square). It will be appreciated that these cards represent both combinations of values of the antecedent and consequent: P(Q), P($\bar{Q}$), $\bar{P}$(Q), $\bar{P}$($\bar{Q}$), Q(P), Q($\bar{P}$), $\bar{Q}$(P), $\bar{Q}$($\bar{P}$).

Procedure. The subjects were allocated alternately to the groups and tested individually. Six subjects in the experimental group were tested with sentence (1) and seven with sentence (2), these proportions being reversed in the control group. Before presenting the conditional sentence the subjects in both groups examined the cards briefly to familiarize them with the fact that there was always a geometric shape on one side and a coloured scribble on the other side.

In the experimental group the conditional sentence was presented and the four programme cards were handed to the subjects who were asked to pick out "the one card which makes the rule false" (i.e. falsifies the conditional sentence). They were then asked to pick out any which "prove the rule true." It was explained to them that their decisions meant that the converse of the sentence could not be assumed—"that the rule only held one way." The subjects in the control group were given a similar amount of time to understand the conditional sentence without any explanation of its meaning.

The selection task had three phases which were the same for both groups.

Phase 1. The eight selection task cards were placed on the desk front side up in a random array, and the subject was asked to pick out "all those cards, but only those cards, which would show you, if you knew what was on the other side, that the rule was true or false."

Phase 2. The subject was asked to turn over all those cards which he had selected and evaluate the conditional sentence with respect to each: "tell me whether each proves the rule true or proves it false."

Phase 3. The subject was invited to project falsity on to the residual cards i.e. those cards which had not been selected. Each pair of cards (representing the same value) was pointed to in turn, starting with the $\bar{Q}$ cards, if these had not been selected: "could anything on the back of those cards make the rule false?"

Results

In the programming in the experimental group all the subjects picked out the P$\bar{Q}$ card without hesitation as the only falsifying instance, and they all picked out PQ as the only verifying instance.

Table IV is the frequency distribution of the choices made in the selection task for both groups. (All subjects were consistent in picking out both instances of any value which they selected).

It is at once apparent that there is little difference in the results of the two groups and that the treatment given to the experimental group failed in its effects.

Table V is the frequency distribution of the evaluation of the contingencies and the projections of falsity on to the residual values in the combined groups. An empty cell in the evaluation task means that at least one of the values associated with that

TABLE IV

FREQUENCIES OF VALUES SELECTED IN BOTH GROUPS

	Experimental group	*Control group*	*N*
P Q	6	6	12
P	5	4	9
P $\bar{Q}$	1	1	2
P $\bar{P}$	1	1	2
P $\bar{P}$ Q $\bar{Q}$	0	1	1
			26

TABLE V

FREQUENCY OF THE EVALUATIONS OF THE CONTINGENCIES AND OF THE PROJECTION OF FALSITY ON RESIDUAL VALUES IN THE COMBINED GROUPS

Evaluation task				*Falsification task*			
PQ	$\bar{P}$Q	P$\bar{Q}$	$\bar{P}\bar{Q}$	$\bar{P}$	Q	$\bar{Q}$	N
t		f		—	—	P	6
t	f	f		—		—	5
t	f	f		—		P	3
t	i	f		—		P	2
t		f	i	—	—		2
t	f	f		Q		P	2
t	f	f	t		$\bar{P}$	P	1
t		f		—	—	—	1
t		f		—	$\bar{P}$	P	1
t	i	f	i		—	P	1
t		f		Q	$\bar{P}$	P	1
t	f	f	i				1
							26

t = true, f = false, i = irrelevant. Note that no P column is included under falsification because all subjects selected P.

cell was absent from both sides of a selected card. An empty cell in the falsification task, however, means that a value on the front of a card, associated with that cell, had been selected (and evaluated), and hence was no longer available for the projection of falsity. A dash in a cell means that a subject denied that any value, associated with the value in that cell, would falsify the rule, and an entry of a value in a cell means that it, associated with the value in that cell, would falsify the rule. It will be noted that the selected values, other than P which was always selected, correspond to the *empty* cells in the falsification task.

It is of particular interest to note that six out of a possible 17 subjects failed to say that P, projected on to $\bar{Q}$, would falsify the rule; and that eight of the 10 subjects, who had claimed that $Q(\bar{P})$ falsified the rule, subsequently said that no value associated with $\bar{P}$ would falsify the rule.

Only nine of the 24 subjects, who had not selected $\bar{Q}$, said that they were aware of having made a mistake in the selection task.

General Discussion

The results show that two kinds of therapy do not facilitate the act of making the contrapositive inference, the selection of $\bar{Q}$, but they do show phenomena of considerable interest. The selection task was not meaningless to the subjects. Their results are far from random. In the combined experiments 50 per cent. initially select just P and Q out of the possible 15 combinations of values which could have been selected. This marked tendency to pick out only those values which are mentioned in the conditional sentence suggests that the selection task seemed deceptively easy. Its real meaning, the challenge which it implies, escaped the subjects to a large extent. But the selection of P and Q, which is so resistant to therapy, is consistent with the theory that individuals are biased through a long learning process to seek and expect a simple correspondence to hold between sentences and states of affairs. The introspections corroborate the theory. "You first of all accept all the cards as true—you don't make any allowances for any of them being wrong until you turn them over." "A rule is a rule, so looking at it frankly the ones with squares will have red scribbles on the back." "I feel very unhappy about my original choice, but yes, I would still choose the same ones if I had to do the task again." One subject twice projected truth instead of falsity before he could be prevailed upon to comply with the instructions.

It is a reasonable inference that this set for truth inhibits the perception of a card as being an exemplar of $\bar{Q}$. Even when it is recognized as such it is not always used to make a deduction. In the results of the combined experiments 16·7 per cent. of the subjects select $\bar{Q}$. But of those who do not initially select it, 30·6 per cent. fail to project P on to it as a means of falsifying the conditional sentence. This seems extraordinarily capricious. It is as if someone had said that any number other than 3 on the other side of D would falsify the conditional and in the next breath denied that any letter on the other side of 7 would do so. However, once the drift of the subjects' reasoning is apparent the logical discrepancies begin to look more plausible. $P(Q)$ and $Q(P)$ come to the same thing when the cards are turned over: they both verify the rule. And when $Q(\bar{P})$ is turned over it certainly doesn't verify the rule—hence it must falsify it. But of course there is nothing on the other side of $\bar{P}$ which would falsify the rule because there is nothing on the other side which could verify it. One might risk a general statement about this. Suppose that a rule is confirmed whenever one state of affairs (y) depends upon another (x); and suppose initially one has access to either x or y but not both. If y is obtained,

then there is an expectancy that x will be found to have been associated with it, thus confirming the rule; but if x is found not to have been associated with y, we are tempted to say that the rule has been refuted. However, if initially x is not obtained, no such expectancy is generated, and we are less inclined to say that, if y were to follow, the rule would be refuted. And if y is not obtained initially nothing seems to follow about the rule at all.

The results, however, are still disquieting. If Piaget is right (Inhelder and Piaget, 1958), then the subjects in the present investigation should have reached the stage of formal operations. A person who is thinking in these terms will take account of the possible and the hypothetical by formulating propositions about them. He will be able to isolate the variables in a problem and subject them to a combinatorial analysis. But this is exactly what the subjects in the present experiment singularly fail to do. The variables in the present tasks are abstract but they are distinct and susceptible to symbolic manipulation. Could it then be that the stage of formal operations is not completely achieved at adolescence, even among intelligent individuals?

It may not, however, be the concept of implication which causes difficulty so much as its customary verbal guise: "if P then Q." That form of words suggests that the consequent follows the antecedent in time, or even that there is a causal relation between them. Other expressions have been tested. Hughes (1966) has shown that the logically equivalent expression, "Q if P," causes, if anything, even more difficulty. But implication can be formulated with simpler logical connectives, e.g. "$\bar{P}$ or Q" (either not-P or Q), "$-(P\bar{Q})$" (it is not the case that P and not-Q). P. N. Johnson-Laird (personal communication) has reported that the mode of formulation makes a considerable difference when the contingencies have to be evaluated. We intend to test these and other formulations to see whether similar differences can be detected in selection tasks. This problem may be generalized. Is it the formal structure of rules which is responsible for their difficulty, or is it the words with which we express these rules? And if the latter, what words illuminate the structure?

I am particularly indebted to my colleague, Dr. P. N. Johnson-Laird, for an invaluable criticism of earlier drafts of this paper. I should also like to thank Dr. Margaret Donaldson and Mr. Roger Wales, of the University of Edinburgh, for kindly organizing the subjects and generously giving me facilities for conducting Experiment II. Finally, I owe a debt to my subjects. Their interest and enthusiasm has been a source of constant gratification.

References

HUGHES, M. A. M. (1966). *The Use of Negative Information in Concept Attainment.* Unpublished University of London Ph.D. thesis.

INHELDER, B., and PIAGET, J. (1958). *The Growth of Logical Thinking.* New York: Basic Books.

WASON, P.C. (1959). The processing of positive and negative information. *Quart. J. exp. Psychol.*, **11**, 92–107.

WASON, P. C. (1961). Response to affirmative and negative binary statements. *Brit. J. Psychol.*, **52**, 133–42.

WASON, P. C., and JONES, S. (1963). Negatives: denotation and connotation. *Brit. J. Psychol.*, **54**, 299–307.

WASON, P. C. (1964). The effect of self-contradiction on fallacious reasoning. *Quart. J. exp. Psychol.*, **16**, 30–4.

WASON, P. C. (1966). Reasoning. In Foss, B. (Ed.), *New Horizons in Psychology.* Harmondsworth: Penguin Books.

Manuscript received 2nd April, 1968.