1979

# THE ROLE OF CONTENT IN REASONING

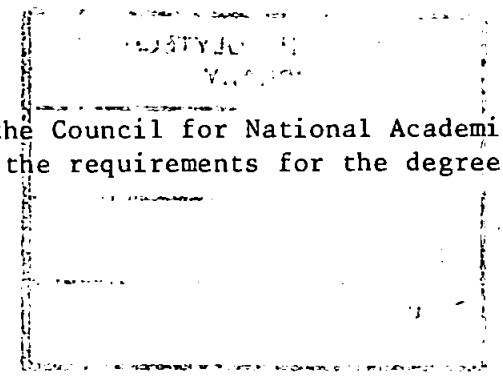
MANKTELOW, KENNETH IAN

# THE ROLE OF CONTENT IN REASONING

by

KENNETH IAN MANKTELOW

Plymouth Polytechnic November 1979



Submitted to the Council for National Academic Awards in partial fulfilment of the requirements for the degree of Doctor of Philosophy

## DECLARATIONS

At no time during the period of registration for the degree of Ph.D. has the author been registered for any other CNAA or University award.

None of the material contained herein has been used in any other submission for an academic award.

Chapter 5 forms the basis of a paper which has been accepted for publication in the British Journal of Psychology.

A programme of advanced study was undertaken in partial fulfilment of the requirements comprising attendance at a special option course (part of Plymouth Polytechnic BA Psychology course Year III) on Speech, Language and Thought; guided reading in the area of deductive reasoning, under the supervision of Dr J Evans; attendance at relevant conferences.

# A C K N O W L E D G E M E N T S

Grateful thanks are principally due to Dr Jonathan Evans, supervisor par excellence.

I am indebted to the Science Research Council for financial support during the time of the research programme. Without these two parties, this project would never have started.

Dr Evans was ably assisted by Dr Ian Dennis, in his capacity as second supervisor; Dr Dennis and Mr Jim Birrell were particularly helpful in the area of statistics and computing.

Ms Gill Snawdon and Mr Ron Garbett showed great patience and considerable expertise in dealing with my cries for technical assistance, and Ms Carol Gordon has done a splendid job of work on the typewriter. Without these parties, the project would never have finished.

# A B S T R A C T

The role of content in reasoning, by K I Manktelow

A programme of research is reported in which the effects of different contents in two deductive reasoning paradigms were investigated.

A review of the literature showed that the two main determinants of performance are the logical structure of the task, and non-logical performance variables such as 'matching bias'. Matching is a prime determinant of behaviour in abstract tasks, e.g. Wason's Selection task. This is shown by systematically negating logical rule components. However, a large literature indicates that logical performance is facilitated by using thematic materials.

In Experiment 1 these procedures were combined to test competing predictions about their interaction. Under both abstract and thematic materials, performance was as previously found with abstracts - there was no facilitation by thematic materials.

Experiments 2 - 5 investigated possible factors behind this unusual result: it remained unchanged throughout. Discussion of these findings, including a re-examination of previous papers, concluded that thematic materials only facilitated logical performance in conjunction with other helpful contingencies.

Experiment 6 used a truth-table task, unsuccessfully, to pretrain a logically appropriate selection set. However, the truth-table task offered an alternative, logically comparable paradigm for a further inquiry into content effects. Experiments 7 and 8 involved using thematic materials in this tasks along with negated rules. There was significant evidence for a content effect, but not manifested in greater logical performance. The results showed clear evaluation patterns in the thematic task compared with a less definite performance in the abstract task. A theoretical analysis suggested that this effect was due to subjects' re-interpreting the rules rather than 'matching'. Implications of this explanation for a general view of reasoning performance and competence were discussed.

# THE ROLE OF CONTENT IN REASONING

K I Manktelow

## PART ONE: REVIEW AND INTRODUCTION

### CHAPTER 1



Tables

The psychology of deduction has had a history of irregular and increasing growth, which in some ways echoes the development of cognitive psychology as a whole: a restricted scope of research on particular questions in the 1930s and 1940s, a gradual re-emergence and diversification of approach in the late 1950s, followed by something of an explosion of interest, both in terms of approaches and output, in the late 1960s - 1970s. Of the two main fields of deductive reasoning - relational and propositional - relational reasoning was the first to receive attention, with the pioneering work of Stoerring in the early 20th century and the work of Sells and his colleagues in the '30s leading the way. Propositional reasoning attracted attention around 1960, largely because of the impact of the theories of Piaget and the new work emanating from Wason and his co-workers. Both of these approaches are relevant to the research to be reported here, but at this point it will be productive to consider the earlier syllogism experiments, since findings which emerged from them anticipate and parallel the central aspects of the work reported both in this review and in the succeeding experiments. Syllogistic reasoning has been found to be influenced by certain extra-logical factors, in particular the context and content of the problems and non-logical response biases. It was this work which engendered doubts in experimental psychology about the propositional calculus of formal logic as a model for human thought, doubts which will be alluded to and reinforced throughout this review and later in the discussion. (The formal logical bases of the problems will be introduced as this review progresses and as the problems are encountered). Thus a brief consideration of the literature on syllogisms will provide a framework on which to establish the bases both of the experimental work reported

here and the background to it.

A syllogism is a quantified deductive argument consisting of two premises and a conclusion. In a typical syllogistic reasoning experiment, the subject is presented with the premises and either constructs a conclusion or evaluates the validity of a given conclusion. The quantified premises are usually in the form of a universal statement, e.g. 'All A are B' or a particular statement, e.g. 'Some A are B'; these two statements can be affirmative or negative, and so there are four basic premise forms. These are shown in Table 1a. The four premise forms can be put into any 2-way combination, and the pattern of this combination is known as the mood of the syllogism. The form of the argument also varies according to the order in which the terms occur, and formally there are four possible forms, or figures as they are called. Each conclusion has a subject (S) and predicate (P), and these are connected in the premises by a middle term (M). The eight possible permutations of these terms are shown in Table 1b; although in formal logic the subject of the conclusion must occur in the second premise, there is no psychological reason for this restriction (Wason & Johnson-Laird, 1972) and so there are four possible correspondents to the traditional figures, with premise orders reversed to make up the others. To illustrate the central points to be made about deductive reasoning in this introduction, here are two syllogisms in the AEE mood. and third figure:

| | | |
|---|---|---|
| | Premises | All X are Y |
| (1) | | Some X are Z |
| | Conclusion | Some Z are Y. |

This conclusion is obviously valid.

TABLE 1 The structure of the syllogism

(a) Premise forms

| | Universal | | Particular | |
|---|---|---|---|---|
| | Affirmative | Negative | Affirmative | Negative |
| | All A are B | No A are B | Some A are B | Some A are not B |
| Notation: | A | E | I | O |

(b) Figure. The four 'traditional' figures are given in the first row.

S = Subject P = Predicate M = Middle term

| | Figure 1 | Figure 2 | Figure 3 | Figure 4 |
|---|---|---|---|---|
| Premises | M - P | P - M | M - P | P - M |
| | S - M | S - M | M - S | M - S |
| Conclusion | S - P | S - P | S - P | S - P |
| Premises | S - M | S - M | M - S | M - S |
| | M - P | P - M | M - P | P - M |
| Conclusion | S - P | S - P | S - P | S - P |

| | | |
|---|---|---|
| | Premises | All priests are good men |
| (2) | | Some priests are Nazis |
| | Conclusion | Some Nazis are good men. |

As syllogism No. 2 has the same structure as the abstract syllogism No. 1, it is also a valid argument, but the time to judge the validity of the two syllogisms may be different between the two examples, as indeed may be the direction of the judgement itself. This is because they differ in two important respects: No. 2 is composed of realistic sentences, which have been claimed to make these problems easier, and the sentences carry a certain meaning which again may influence reasoning. Plausibility of both premises and conclusion, in this case the relation of priests, Nazis, and the goodness of the two parties, may obstruct logical judgement, so it is not certain whether the realistic content of the examples should make the problem easier or harder. A third factor, one which is not obvious from the syllogisms above, has also been suggested to influence performance: the non-logical response bias of 'atmosphere', which in this case would predict that, since at least one of the premises contains 'some', there will be a tendency for the conclusion also to do so, irrespective of the logical effects of this choice. As the effects of context, content, and response biases all have a well-documented history in the research literature, some more detailed examination of them will be undertaken before proceeding further.

Response biases

The 'atmosphere effect' was one of the first psychological factors proposed to account for observed response patterns in syllogistic reasoning. It has a strong form and a weak form. The strong form, put forward by Woodworth & Sells (1935), states that the terms of the

premises of a syllogism create an 'atmosphere' which pervades the conclusion, such that (i) universal and particular premises lead to universal and particular conclusions respectively, and (ii) affirmative and negative premises lead to affirmative and negative conclusions respectively. Negatives and particulars have a dominant effect, so that if the two premises have at least one of these, the conclusion will be biased in that direction. This predicts that if one premise is negative and one particular, the conclusion should be a particular negative - even though this specific form is not represented in either premise. The weak form of the atmosphere hypothesis arose when Sells (1936) added the 'principle of caution': that subjects are predisposed to accept weak rather than strong conclusions, 'some' rather than 'all'.

Atmosphere was revived and attacked by Chapman & Chapman (1959). They point out that Sells' confirmation of the principle of caution could be artifactual, since the universal logically entails the particular - if 'All A are B' is true, then 'Some A are B' must also be true - so that one would expect more particular conclusions a priori. They criticise Sells on other counts: confusing the mood and figure of his syllogisms and giving the subjects only one conclusion to evaluate with each premise pair. Using a multiple-choice test in which subjects selected a conclusion from the five possible alternatives, including 'no conclusion possible', they found results which conflicted with the predictions of Atmosphere on some syllogisms. They suggest that subjects are using a conversion strategy, accepting the converse of the premise as also true; this is legitimate only for particular affirmatives and universal negatives. Thus errors are due to an understanding of the premises which differs from that dictated

by formal logic. The Chapman & Chapman results, it should be noted, only pose problems for Atmosphere as originally formulated (without the principle of caution) on two premise types, indicating perhaps that a combination of atmosphere and conversion might best account for the data, a position adopted by Begg & Denny (1969). Frase (1966) has also suggested that the predictions of Atmosphere may be confounded with the logical definition of 'some'. Other workers have argued that errors stem from faulty (Henle, 1962) or inadequate (Ceraso & Provitera, 1971) analysis of the premises.

More recently, Johnson-Laird (1975) has pointed out a different kind of atmosphere effect, which he calls the 'figural' effect, resulting from the order in which the terms appear in the premises, and which seems to operate when the terms 'cross over'. Thus, given the syllogism A-B, B-C, where A-B indicates the order in which the terms are mentioned in the premises, 85% of subjects gave a conclusion in the form A-C, whereas given B-A, C-B, 86% of the conclusions were C-A. There was little evidence of such biases when the connecting term B was either mentioned first in both or second in both.

One can readily appreciate from this brief examination that the literature on response biases in syllogistic reasoning is by no means in total accord, and that any account of reasoning based solely on Atmosphere or figural effects is inadequate. For instance, such theories say nothing about how subjects arrive at correct deductions or 'no conclusion possible' answers, both of which account for large proportions of the responses in studies where they are both available. Neither do they specify what leads subjects to succumb to these tendencies in the first place. The point is however that the idea of

response biases which cut across logical reasoning processes was established by these studies, and that even if they do not provide a wholly satisfactory account of the observed behaviour, there is reason to believe that they have some influence. In propositional reasoning, as we shall see, non-logical response biases have a much stronger claim for acceptance.

Content and context

From response biases we move on to variables inherent in the materials which make up the premises of these arguments: content and context effects. These are closely linked: the context of an argument will obviously be reflected in the content, and similarly a particular content, especially if in thematic terms, will exist in and involve some context, but broadly these break down in the research literature into two variables - realism of materials (content) and the effects of prior beliefs (context).

The belief-bias effect is a regular feature in the syllogistic reasoning literature; it appears in the early history of the research (Wilkins, 1928), and is still going strong (Revlin & Leirer, 1978), with a fairly even scatter over the intervening 50 years. Briefly, the effect is that when logic and belief conflict, logical accuracy deteriorates - people tend to accept conclusions which fit their beliefs, irrespective of logical validity. One of the classic experiments was that of Janis & Frick (1943), who balanced validity and invalidity of conclusions and agreement/disagreement with subjects' beliefs, the latter being assessed by an attitude test. They found a tendency for subjects to accept an invalid conclusion which agreed with their beliefs, and to reject valid conclusions which did not. This was a marginal effect, since only 23% of all judgements were erroneous, and a third

of these errors did not follow the belief-bias predictions. Morgan & Morgan (1953) found that logical performance on similar types of problems to those used by Janis & Frick was improved by three hours' training in formal syllogistic logic before testing: Frase (1966) later established that this was due more to an effect on the appreciation of the logical quantifier than any diminution of the belief-bias effect per se. These are only a few examples of the kind of study done, but they serve to illustrate the belief-bias effect, a consistent though theoretically troublesome finding (Morgan & Morton, 1944; Henle, 1962).

As for the simple effect of different types of content on reasoning performance, the most frequently reported is that of the difference between abstract (or symbolic) and thematic (or concrete, meaningful, realistic, familiar) materials. It is commonly stated in the literature that reasoning, both syllogistic and propositional, is more logical when problems contain thematic materials than when they are in abstract form (see examples (1) and (2) above). We shall return to this question as it affects conditional reasoning later; looking at syllogistic reasoning specifically, it is in fact difficult to find studies devoted to this particular topic. There are many which use abstract materials and many which use thematics, but few which compare the two. Sells (1936) reports a thematic materials effect, and Wason & Johnson-Laird (1972) qualify the discussion of Atmosphere by restricting it to abstract materials (p. 133), but the study which has become the cornerstone of the argument is that of Wilkins (1928). It is mentioned throughout the syllogistic reasoning literature, and cited in several studies of content effects in conditional reasoning (e.g. Wason & Shapiro, 1971; Johnson-Laird, Legrenzi & Legrenzi, 1972; Van Duyne, 1974; Staudenmayer, 1975). It is appropriate therefore to

examine this paper more closely, in the light of its undoubted historical significance.

Wilkins gave 81 college students 160 syllogisms comprising 40 problems in 4 types of materials. There were three conclusions after each syllogism and the subjects had to evaluate the validity of each, so they had to make 480 evaluations. This test was spread over three hours in two separate sessions. Three of the problems were in fact transitive inference tasks and not quantified syllogistic arguments of the type under discussion. The scoring system was to compute for each subject the percentage of correct evaluations for the items attempted - an accuracy score controlling for, but not reflecting, speed. Wilkins' analysis is based on three parameters: mean accuracy scores, correlations between material types and between syllogism scores and intelligence tests, and inspection of individual data. Taking these in order: for the two types of content which most closely correspond to those used in the papers and experiments to be reported, percent correct responses are - Thematic: 84.6%, Abstract: 75.6%. The correlation between scores with these materials is +.70 (N=80). On inspection, it seemed that more subjects found greater difficulty with abstract materials than with thematics, the difference in accuracy scores between materials being greater for the former set of subjects. There is no statistical analysis to establish whether these differences are significant, in fact the correlation seems to indicate that, as Wilkins puts it, "there is a high degree of relation between ability to reason with familiar material and ability to reason with more abstract material". Wilkins makes much of individual differences and emphasises how wide they are, and his further conclusions reflect this. To quote again: "it would seem that changing the material does to some

extent change the position of some individuals in regard to their ability to do (syllogistic) reasoning. That is, some individuals do better with more abstract material than with more familiar and concrete material; and others do better with familiar material."

Some points need to be made here, since they will recur later. Firstly, there is little evidence to suggest an overall facilitation of reasoning by realism in this study, in spite of a somewhat inconsistent conclusion to that effect in Wilkins' own summary. Such differences as there are are not of the order claimed in some propositional reasoning papers - a score of 75% logically correct with abstract materials would be regarded as freakishly high in conditional reasoning research - and seem at best to be marginal tendencies in some individuals. It is therefore apparent that to presume, without qualification, that thematic materials facilitate logical reasoning is to run away with an unwarranted conclusion, supported only by data from a single, old, statistically inconclusive study. This is a theme which is to reappear after consideration of the literature on propositional reasoning, and it is to this which we now turn.

## CHAPTER 2

In the review and experimental reports which follow, two forms of propositional inference are examined: conditional ('If... then...' statements) and disjunctive ('Either...or...' statements). There are other forms, but these two are much the most extensively investigated. Between them, conditional reasoning has received by far the most attention, perhaps because the difference between formal logic and actual behaviour is much more apparent, and elusive of explanation, for conditionals than disjunctives. It is only fairly recently that propositional reasoning as such has been studied, and that study has involved three main paradigms, which though distinct are closely related, and which may be called (i) Inference tasks, (ii) Truth-table tasks, and (iii) the Selection task. All these paradigms can be, and have been, used to explore both conditional and disjunctive reasoning. (Conditional and disjunctive sentences have been used in other fields of inquiry, e.g. concept attainment, but these fall outside the scope of the present dissertation).

Problem structure

Inference tasks are similar to syllogistic tasks in that they use two-premise deductive arguments leading to a conclusion which may be valid or invalid. However, they differ in their internal structure; conditional inference tasks - disjunctives will be discussed later - are said to involve the logic of material implication. This 'said to' is important, since it has become clear through the progress of research that to invoke any one formal system as a standard against which to measure people's performance may be a mistake, as other logical systems and systems with no formal standing may be more efficient at describing the data. However, material implication has

often been cast in the role of a competence model for conditional reasoning, and has a long history in this guise, and so an examination of its formal logic is an appropriate point at which to open on inference tasks.

A conditional argument involves a rule of the form 'If p then q'. This forms the 'major premise' of the argument. The next step, the 'minor premise', is a statement of the antecedent (p) or consequent (q) in affirmed or negated form, from which follows, validly or fallaciously, the conclusion. There are thus four possible permutations of minor premise and conclusion, each having been given a specific name. The structure and notation of these arguments appear in Table 2. It is an axiom of material implication that only Modus Ponens and Modus Tollens are valid inferences, Denying the Antecedent and Affirming the Consequent being fallacies; in material implication, the conditional does not imply its converse. This can be seen in an example such as 'If it is a tiger, then it has stripes'; there are many other striped things, animate and inanimate, which are not tigers. Following the four inferences then we can see that, given that something is a tiger, we can justly conclude that it is striped (MP); fortunately, given stripes we do not have to conclude 'tiger' (AC); that something is not a tiger does not mean that it cannot have stripes (DA); but it is quite valid to conclude that if something has no stripes, it cannot be a tiger (MT). The neatness of this system is however disturbed on two counts: firstly, there are some conditionals which clearly do imply their converses, and therefore fit an alternative logical system (material equivalence; see Table 2); secondly, even given an implication conditional, the four rules of inference do not necessarily reflect how people actually reason from

TABLE 2 Conditional and disjunctive inferences. Valid (V) and invalid (I) references for implication and equivalence conditionals, and inclusive and exclusive disjunctives, are shown, with the names by which the various inferences are commonly known.

Conditionals

Major premise: 'If p then q'

| Minor premise | Conclusion | | Implication | Equivalence |
|---|---|---|---|---|
| p | q | Modus Ponens (MP) | V | V |
| not p | not q | Denying the Antecedent (DA) | I | V |
| q | p | Affirming the Consequent (AC) | I | V |
| not q | not p | Modus Tollens (MT) | V | V |

Disjunctives

Major premise: 'Either p or q'

| Minor premise | Conclusion | | Inclusive | Exclusive |
|---|---|---|---|---|
| p | not q | Affirming the First Component | I | V |
| not p | q | Denying the First Component | V | V |
| q | not p | Affirming the Second Component | I | V |
| not q | p | Denying the Second Component | V | V |

it. It is by no means easy to decide, beyond stating a few general qualifications, when a conditional rule should be assumed to be one of implication or equivalence. In logic, the issue is clarified by stating the rule of equivalence as 'If and only if p then q', but this rather wordy form is uncommon in natural language, and the issue is usually decided by semantics. Rules of equivalence tend to be binary statements or generalisations (Wason & Johnson-Laird, 1972), definitions, causal connections, threats, or promises - and this is not an exhaustive list. Researchers have attempted to get round this semantic problem by using abstract materials, but as we shall see, this produces problems of another kind.

Abstract materials are the basic tools of the inference task trade; they were used at the outset of research and continue to be used. Their use is intended to obviate extraneous biases resulting from the plausibility of conclusions to meaningful sentences, such as the belief-bias effect noted in the syllogism literature, and the interpretation biases for certain conditionals alluded to above. With abstract materials it is difficult to see how conclusions could be deflected by a person's beliefs, or how that person could justifiably interpret a rule of material implication as one of equivalence. The favoured form of abstract content is letter and number pairs, e.g. 'If the letter is L, then the number is 5' (Evans, 1977a); sometimes only letters are used (Roberge, 1971a, b, 1974, 1978). In fact, the manipulation of problem content, in the form of variations of both the syntactic and semantic forms of rules, has occurred as something of an appendage to the research with the basic abstract conditional, as if the latter provides some idea of basic performance which is qualified by considerations of content and context. This is a view underlying truth-table and Selection task research as well, and one

which will not receive wholehearted endorsement in the ensuing discussion. However, it is a fact that abstract materials have been the most popular, and so it is appropriate to survey the empirical findings beginning with them.

Experiments: basic evidence

Having thus set out the problem, it may be mildly surprising to find that the population of published studies is not large, and rather more surprising to note the dearth of published experiments on the plain inference task as outlined above: an 'If p then q' abstract rule and the four inferences. Logicians (e.g. Strawson, 1952) have appreciated for a long time that error - in the sense of failing to adhere to the formal calculus - was common on the DA, AC, and MT inferences, and rare on the MP inference. Do the empirical findings bear this out?

For a proper examination of subjects' performance on the inference task one has to collate data from more complex experiments, where the 'standard' task is embedded in a multi-factorial design, or from experiments which only look at particular inferences. On doing this, it is immediately apparent that there is a great deal of variability between the studies regarding the frequency with which subjects make the various inferences, and this makes the development of a precise theoretical account of inference task performance a risky business, since only broad conclusions are possible.

Wason & Johnson-Laird (1972) cite data from a study by Hill, reported in Suppes (1965), which show that children possess a high degree of logical competence; however, from the results of an experiment by Shapiro which they also cite it appears that one cannot generalise from developmental studies to those on adults, since

Shapiro's data seem to tell a different story. In a task in which subjects had to evaluate the validity of all four conditional inferences, using abstract materials, the frequency with which each inference was judged as valid was as follows: MP: 95%, DA: 25%, AC: 20%, and MT: 48%. Evidently MP was a basic inference; and there was a strong tendency not to fall prey to the fallacies, although there was still a considerable number made. The most interesting finding was that only around half of the MT inferences were considered valid. An examination of the possible reasons for this and other findings will proceed following an account of some other published experiments, most of which were conducted after the Shapiro study.

Results from these experiments,all using abstract materials, are presented in Table 3: some points should first be noted in inspecting these data. Firstly, not all the inferences are equally represented - MT and AC have received the greatest attention, as might be expected. Secondly, in only one experiment were subjects allowed to construct their own inferences from the premises (Roberge, 1978); among the rest, subjects had either to evaluate the validity of a given conclusion, as in the Shapiro study, or choose a conclusion from a set of alternatives. The experiments are listed in two groups on this basis. In the evaluation procedure, some subjects were given a two-way choice between valid and invalid, others a third choice of 'maybe' or 'indeterminate', and these are denoted by a (2) or a (3) in the Table. In the Shapiro study, Wason & Johnson-Laird report that the task was "to decide whether or not the inference was valid", and we therefore presume that there was no 'indeterminate' alternative. Thirdly, a number of important investigations are not represented in the Table (e.g. Taplin, 1971, Taplin & Staudenmayer, 1973, Staudenmayer, 1975, Rips & Marcus, 1977). This is not because

of any anti-American feeling, but because they do not present frequency data in a usable form. They do, however, have a singular importance of their own, and will not be denied a hearing.

Turning to the data in Table 3 then: the standing of MP as a basic pattern of inference is confirmed, with only a tiny minority dissenting. In the Roberge (1971a) results, the only abstainers were those who answered 'maybe' to the MP inference - no-one denied it. There is little else to say about the MP results here; evidently, MP expresses the very meaning of the 'If...then...' conditional. The same situation does not recur with the other inferences, although the overall pattern seems to be that there are fewer DA inferences, made than AC and MT; in only one of these findings does the frequency of DA exceed AC or MT. It is also plain that there is little evidence for any behavioural difference between selection and evaluation procedures, as the means show, despite a conclusion by Evans (1972a) to the contrary. The only obvious difference seems to be in AC frequencies, and here it appears that the data from the Evans (1972a) study have elevated the evaluation mean score. Indeed, the data from this study are generally out of line with others', and one feels obliged to look for reasons. In doing so, the points made will serve not as a critique of Evans' experiments but rather as an indication of the susceptibility of inference tasks to 'slight' procedural changes - the changes are certainly not slight in their effects - and of the need to be guarded on this count in assessing the conclusions and explanations forthcoming. The data do, in fact, come from a study in which three other conditionals, containing negative rule components, were also given to the subjects, so perhaps this complication of the tasks brought about the unusual results. However, Roberge (1971a)

TABLE 3 Percentage frequencies of the 4 conditional inferences given in the literature on the 'basic' task: affirmative rules and abstract materials only. See text for notation

| | | Inferences | | | | |
|---|---|---|---|---|---|---|
| Evaluation Experiments | Sub-conditions | MP | DA | AC | MT | N |
| Shapiro (unpub.) | (2?) | 95 | 25 | 20 | 48 | 20 |
| Roberge (1971a) | (3) | 97 | 28 | 45 | 45 | 110 |
| Evans (1977a) | (2) | 100 | 69 | 75 | 75 | 16 |
| Evans (1972a, Expt II) | (3) | | | 90 | 71 | 16 |
| Weighted Mean | | 97 | 32 | 49 | 51 | |
| Selection Experiments | | | | | | |
| Cope (1979, Expt I) | Binary | 100 | 45 | 65 | 55 | 54 |
| | Non-binary | 94 | 22 | 50 | 22 | |
| (Expt II) | Binary | | | 57 | 79 | 54 |
| | Non-binary | | | 44 | 40 | |
| Evans (1972a, Expt I) | Non-binary | | | 32 | 91 | 16 |
| Roberge (1978) | Selection and construction | | | | 70 | 64 |
| Weighted Mean | | 97 | 34 | 52 | 56 | |

used an even more complex design, involving negatives, logical falsehoods, and transitive inferences, so this is unlikely. It is possible that in looking at only two inference types, the demand characteristics of the experiment were changed slightly, and this receives some support from the data from the selection form of the task, at least on MT, where frequencies are generally higher for procedures which involve only one or two inferences, though it should be noted that only one study provides the data for the procedure of using all four inferences, that of Cope (1979). There seems to be no systematic pattern associated with giving the subjects two or three choices in the evaluation task or variations in the array of choices in the selection form of the task. The most likely reason is that Evans gave his subjects, in his first experiment, a pretest using AC and MT, and told them whether or not they were right.

Cope (1979) investigated a hypothesis, derived from results of a truth-table experiment (see Chapter 3) by Legrenzi (1970), that the tendency for subjects to interpret a rule of implication as one of equivalence may be increased either by use of the rule as a causative statement, or by materials being of a strictly binary nature, these factors being compounded in Legrenzi's experiment. Cope uses abstract letter-number materials, which prevent a causal connotation, and defines these as coming from populations consisting of either just two letters and two numbers (binary condition) or more than two letters and numbers (non-binary condition). He concludes that there is no increase due to the binary presentation in any tendency to treat the 'if p then q' rule as an equivalence, i.e. where the rule can validly be expressed as 'If q then p', and the AC and DA inferences therefore are equally valid. However, his method of analysis is

something of a blunt instrument, in that he only considers in his equivalence category those subjects whose selections are totally in accord with an equivalence interpretation, i.e. those who select all four inferences as valid. Perhaps, since we are dealing with relative tendencies here, he should really be assessing the relative frequency of inferences between conditions, irrespective of the degree of consistency within subjects. These frequencies may be examined in Table 3, where it appears, in the absence of a statistical comparison, that there is indeed a higher frequency of DA, AC, and MT inferences, which would suggest a heightened tendency towards an equivalence interpretation. Cope's contention that it must have been the causal connotation in Legrenzi's experiment which brought about the effect is therefore doubtful. One might also comment that Legrenzi used a truth-table task and Cope an inference task, and propose that since these are two different things, the one is not a proper test of the other. This is not the last appearance of this proposition, and some further discussion of this point in relation to other literature is given in Chapter 3.

The question posed before, and the one reflected in the census of experiments in Table 3, is the one to which we now return: why do subjects apparently find more difficulty in appreciating that MT is as valid as MP? This question has not only attracted considerable attention in itself, it has also led to the posing of some equally challenging additional questions.

Wason & Johnson-Laird (1972) outline several strategies by which a person might come to make an MT inference. He may be said to 'possess' the rule of MT in his repertoire, much in the way people seem to possess MP; failing this, he may learn it by experience, recast

it in a more usable form or an alternative logical equivalent, or operate by a method of deduction known as Reductio ad Absurdum. In doing this the subject starts with a basic hypothesis derived from the rule, e.g. he might say, given the rule 'If p then q': Suppose p; by MP I conclude q. But not-q is stated; this is an absurdity, since p cannot imply both q and not-q. Therefore, to resolve the contradiction, I conclude not-p. Failing all these, the reasoners may even just guess the answer, but as they do not seem to do so on any of the other inferences this is unlikely. Of course, all these proposals are in a sense superfluous, since almost half the MT inferences are incorrectly used. Wason & Johnson-Laird go on to conclude that the most likely source of difficulty is the presence of a negative in the minor premise of the MT argument. There is a good deal of suggestive evidence for this: negatives have been found to bring about increased difficulty, both in terms of speed and accuracy, in a number of paradigms. To appreciate the importance of negation to the study of reasoning, a brief digression on the relevant experiments is in order here; we shall also meet negation again in the discussion.

Negatives

In a series of pioneering experiments, Wason (1959, 1961, 1965) found evidence of an interaction between negation and the truth value of a sentence, reflected both in response times and error frequencies (Wason & Jones, 1963). There are four types of sentence in these experiments, corresponding to the possible permutations of nagation and truth value, and these are listed below together with examples from the materials used in the 1961 experiment.

| Affirmative | | Negative | |
|---|---|---|---|
| True (TA) | False (FA) | True (TN) | False (FN) |
| 24 is an even number | 39 is an even number | 57 is not an even number | 92 is not an even number |

The order of difficulty of evaluating these sentences is surprising:

TA<FA<FN<TN

Thus the effect of the negative differss between the true and false sentences. This seems surprising because the expression 'false negative' conveys a sense of 'double negation', whereas a 'true negative' appears to be only singly negated. However, these connotations disappear when one separates the two kinds of negation which inhere in these sentences: one is the possible semantic mismatch between the sentence and its referent, e.g. the word 'even' and the number 57, and a syntactic mismatch due to the presence of 'not'. These are both present in the TN sentence, but there is only the syntactic negative in the FN; Wason (1972) further clarifies the point by calling the sentences 'denial of a falsehood' and 'denial of a truth' respectively. This kind of analysis has been incorporated into information-processing models of negation, details of which need not detain us here (e.g. Clark & Chase, 1972; Carpenter & Just, 1975), which suppose that the true affirmative is a fundamental linguistic unit and that negation, of whatever type, produces difficulty.

These deliberations about the processing of negatives have so far taken place without considering their role in language - the context in which they would actually be used. Wason (1965) proposed that the difficulty of negation would be diminished if negatives were used in their natural place: "the contexts of plausible denial". This proposal was explored through two hypotheses, briefly stated as

the 'exceptionality hypothesis', that negation constitutes plausible denial when it refers to an exceptional attribute, and the 'ratio hypothesis' that, given two sets of stimuli of differing magnitude, it is more plausible to deny that the smaller set has the characteristic of the larger than vice-versa. The results (response times) confirmed the first hypothesis but not the second, although subsequent experiments (e.g. Cornish, 1971) have found it quite possible to derive contexts for denial on the basis of the ratio hypothesis. Greene (1970) attacked the context question from a different angle: that negation is a matter between two sentences rather than between a sentence and a physical situation, and that the function of negation is therefore to "signal a change in meaning". Negatives are unnatural when used to preserve meaning. Greene constructed pairs of sentences corresponding to these natural and unnatural uses of a negative, and asked her subjects to separate, in a card-sorting task, those pairs which meant the same thing from those whose meanings differed. An example of a natural pair is as follows:

x exceeds y; x does not exceed y

Here it is easy to see that there is a difference in meaning. However, given the unnatural pair.

y exceeds x; x does not exceed y

it is not so easy to appreciate that these mean the same thing. Data from this and other experiments provided strong confirmation of this effect.

It is revealing to note that the natural negatives in Greene's study correspond to the false negatives in Wason's earlier experiments - the denial of a fact. This, as Wason himself has pointed out, emphasises the central function of negatives in natural language:

denying preconceptions. One of Wason's own examples will illustrate this. The negative in the statement 'the train wasn't late this morining' looks like a true negative on its own, but when the reason for the utterance is taken into account - the context of the usual lateness of the train - it is exposed as a false negative. It does not have to be subjected to the various processes described before, of comparing semantic and syntactic mismatches, because it is serving its natural function of denying the preconception inherent in the statement itself.

Inference experiments with negatives

Negation, then, has been found to lead reliably to certain patterns of difficulty, which can be overcome by asserting a context in which it can operate, or by playing it in its natural role as a meaning changer. One strategy for illuminating this process of MT therefore would be to vary the presence of negatives in the deductive argument: if negation is behind the apparent difficulty of MT, then manipulating it should result in concomitant variations in MT reasoning. This has been done by two independent researchers, Roberge and Evans, both of whom have looked at all the four inferences as well as MT in particular. Table 4 shows the frequencies with which the inferences were made in their experiments, on the rules which contain negatives - results from the affirmative rules are included in Table 3. There are four possible types of sentence in these experiments, according to the combinations of negative and affirmative antecedent and consequent: AA (affirmative antecedent and consequent), AN (affirmative antecedent/negative consequent), NA (negative antecedent/affirmative consequent), and NN (negative antecedent and consequent).

There are several tendencies arising from the studies which

TABLE 4 Percentage frequencies of the 4 conditional inferences made from rules containing negative components; data from four studies listed in Table 1.

| | Procedure | Inferences | | | |
|---|---|---|---|---|---|
| AN rule - If p then not q | | MP | DA | AC | MT |
| Roberge (1971a) | Evaluation | 99 | 9 | 53 | 46 |
| Evans (1977a) | Evaluation | 100 | 13 | 31 | 56 |
| Evans (1972a, Expt. I) | Selection | | | 35 | 75 |
| Roberge (1978) | Selection/ construction | | | | 78 |
| Weighted Mean | | 99 | 10 | 48 | 62 |
| NA rule - If not p then q | | | | | |
| Roberge (1971a) | Evaluation | 93 | 26 | 59 | 15 |
| Evans (1977a) | Evaluation | 100 | 50 | 81 | 13 |
| Evans (1972s, Expt. II) | Evaluation | | | 100 | 6 |
| (Expt. I) | Selection | | | 61 | 38 |
| Roberge (1978) | Selection/ construction | | | | 55 |
| Weighted Mean | | 94 | 29 | 66 | 27 |
| NN rule - If not p then not q | | | | | |
| Roberge (1971a) | Evaluation | 97 | 9 | 45 | 30 |
| Evans (1977a) | Evaluation | 100 | 19 | 81 | 25 |
| Evans (1972a, Expt. I) | Selection | | | 55 | 41 |
| Roberge (1978) | Selection/ construction | | | | 66 |
| Weighted Mean | | 97 | 10 | 50 | 42 |

have looked at all four inferences (Roberge, 1971a, b; Evans, 1977a). Evaluation of the two Roberge papers, which almost certainly report the same set of data in different forms, is difficult, since in one form (1971a), a table of response frequencies is presented without further analysis, and in the other (1971b) a statistical analysis based on errors - i.e. responses which depart from the model of material implication - is presented. The situation is further complicated by the inclusion in the design of logical falsehoods, an example of which would be an ordinary MP argument with 'not q' as the conclusion instead of 'q'. This is not in itself a sin, and the falsehoods do assess subjects' readiness to say 'no' as well as 'yes', but as the responses from them do not vary significantly from those on the normal inferences and are grouped with them in the analysis, they serve mostly as an unnecessary supplement to the error term. In addition, the analysis in the 1971b paper is simply in terms of polarity (affirmative/negative) of premises and conclusions, not of inferences made on particular rules. That Roberge found no difference in incidence of errors between affirmative and negative conclusions is not altogether surprising, since 'error' in these terms is exactly balanced across the design according to inference, sentence types, and truth/falsehood; the analysis cannot reflect differences in inference patterns over rules. For this we have to inspect the relevant parts of the 1971a paper (reproduced in Tables 3 and 4) and compare them with an equivalent study using a statistical analysis not based on a prior assumption of formal logic as a competence model. Luckily, such a study does exist: that of Evans (1977a), and there is a reasonable level of agreement, as far as can be judged by inspection, between it and the Roberge results. Roberge (1971b) provides a

foretaste of Evans' analysis in some of his findings: subjects make more errors on the NA rule, and he notes that when an affirmative conclusion has the opposite polarity to its corresponding term in the rule, subjects give more 'maybe' responses. He also notes that there is a noticeable stability of responding on inferences where the minor premise involves affirmation (MP and AC), regardless of the form of the major premise or conclusion.

The Evans (1977a) data echo these findings and enlarge them considerably. As the analysis is not in terms of errors, it is a more straightforward task to examine the inferences. Again, it was found that MP was stable across rules - in fact, all the subjects evaluated the inference correctly on all rules. As for the other inferences, it was found that there were more DA's on rules with affirmative consequents, more AC's on rules with negative antecedents, and more MT's on rules with affirmative antecedents. These results coalesce into what Evans nominates as a generalised response tendency - hinted at by Roberge (1971b) - for subjects to prefer inferences whose conclusion is negative; thus there are more DA's and MT's when denial of an affirmative is possible, and more AC's when affirmation of a negative is possible. It should be noted that this effect, for subjects to prefer negative conclusions, seems entirely absent on the ordinary AA rule: there is only a slightly higher incidence of MT (negative conclusion) than AC (affirmative conclusion), and a much lower incidence of DA (negative conclusion). Negative conclusion bias seems to be a specific product of negation in the rule itself, comparable to the atmosphere effect in syllogistic reasoning. This gives a further clue as to the role of negation in inference making, and to elucidate it we need to go back in time to the study by Evans (1972a), which looked at

MT and DA only.

It was the first experiment in this study which first clarified the possibility of a negative conclusion bias, and in fact all the findings of the 1977a experiment detailed above were predicted from its results. A brief inspection of Table 4 will confirm that the findings of Evans (1972a) Experiment I concur with those of the two experiments outlined above: at this juncture, it is the second experiment which is of most interest. Wason & Johnson-Laird (1972) and Wason (1972) point out that the difficulties due to negation in reasoning, both with MT and negative rules, could be computational or conceptual: there could be processing difficulties in dealing with multiple negations, or a breakdown in the subject's grasp of the meaning of negation in the context of the experiment. Evans (1972a, Experiment II) aimed at separating these. The argument goes thus: in the MT inference with an affirmative-antecedent rule, the correct conclusion involves a direct denial of p, so that not-p follows readily. However, when the antecedent is negated, the conclusion involves the denial of not-p - not not-p - and there seems to be a process of 'double negation' to go through before the correct conclusion (p) is reached. The difficulty could be due, as stated above, to the additional process involved in the 'double negation', or to a failure to appreciate that a nagative which, as Wason (1965, 1972) has emphasised, would normally be used to express falsity, could itself be false. Evans attacked this problem by using logical falsehoods, which we have met before. In an MT argument, the logically warranted conclusion is 'not p', which should be evaluated as true; the falsehood, the conclusion 'p', logically requires the evaluation of false. Thus in an MT argument with an NA rule, evaluating the

opposite of not-p, i.e. p, as true, involves the step of double negation, whereas evaluating a 'not p' conclusion as false simply constitutes a direct denial. The double-negation step only occurs on NA (or NN) rules, so there should be a greater difference in MT frequencies here than on AA rules, where direct denial and single negation only are involved. Evans found no evidence for such an interaction, and concludes that the difficulty of MT with negated-antecedent rules is therefore most likely conceptual in origin: subjects are unwilling to infer that an expression of falsity could itself be false. Before leaving the discussion, Evans voices an important caveat, and one which has been confirmed by subsequent experiments (see Table 4): the response profiles differ between selection and evaluation procedures in these experiments when negatives are involved. The difference is only on the MT inference - AC and DA have not been used in a selection experiment using negative rules - and appears only on the rules containing negatives, both in the antecedent and consequent. On all three such rules, the evaluation MT frequencies are always lower than the selection MT frequencies; the difference is most striking on NA rules. Quite why there should be this suppressive effect on the MT inference when the conclusion has to be evaluated rather than selected or constructed is mysterious, but the effect is consistent and would merit further investigation, since no ready explanation presents itself.

Directionality of the conditional

In general then, there is a wealth of evidence to suggest that negation may indeed contribute to the relative difficulty of the MT inference: negatives generally introduce difficulty into tasks where no deduction is involved, and increase it in tasks where there

are inferences to be made. Negation is unlikely to comprise the whole story though; the MP inference is unaffected by the introduction of negative rules, while MT retains its difficulty even when its inherent negation performs its 'natural' denial function. Perhaps some difficulty may stem from the direction of the inference? After all, MT requires the reasoner to jump backwards from the consequent to the antecedent, so if the jump were made to go forwards, perhaps the inference might be easier. Before looking to the data, an example will illustrate the point, and we return to the MT argument about striped tigers:

If it is a tiger, then it has stripes. It does not have stripes.

Therefore, it is not a tiger.

Would the deduction be made easier if the argument took the following form?

It has stripes, if it is a tiger. It does not have stripes.

Therefore, it is not a tiger.

Note that the rule has not been converted into an equivalence - the universe of striped things is still larger than the universe of tigers, and without the word 'only' before the 'if' it is still the expression of material implication formulated in the above example - one still cannot say, by the AC inference, that because we know it is striped, it must be a tiger.

Braine (1978), in a theoretical paper to be discussed later, emphasises the role of directionality in conditional inferences and cites, regrettably vaguely, some findings in support of the argument above. Evans (1977a), noting the sense of directionality implied by the 'If...then...' form of the conditional, proposed that this could be counteracted by restating material implication in a logically

equivalent form: 'p only if q'. The same inference rules apply, but there seems, intuitively, a different emphasis between two principles of material implication arising from the inference rules but not so far formally stated: the antecedent is sufficient for the consequent, and the consequent is necessary for the antecedent. Returning to the jungle again: given the rule about striped tigers, the presence of a tiger is sufficient, but not necessary, to conclude that stripes are also present; stripes are necessary for there to be a tiger present, but not sufficient to make that conclusion with certainty. The 'If...then...' form and the MP deduction, argued Evans, emphasise the sufficiency of the antecedent, while the '...only if...' form and MT emphasise the necessity of the consequent. There should therefore be more MP inferences made on 'If then' rules than on '...only if...' rules, and more MT inferences made on 'only if' rules than on 'If then' rules. The experiment which tested this also incorporated negated rule components, and when results were pooled across rules the two predictions were both confirmed (see Tables 3 and 4). The result for MT was mostly due to heightened frequencies on rules with negated antecedents. The greater frequency of MT's on 'only if' rules was also found by Braine (1978) and by Roberge (1978), with the difference extending across all rules. It therefore seems that directionality, as well as negation, may indeed contribute to the difficulty of MT, and that by using a form of rule which seems to promote reasoning from the consequent rather than from the antecedent this difficulty can be reduced. However, the 'only if' form may be having this effect for another reason: Evans (1977a) found that there were also more AC inferences judged correct on the 'only if' rules. This leads to a suspicion that the subjects may have been

converting these rules into equivalences. However, the frequency of DA's was the same on both rule-forms and similar to the MT frequency on the 'If then' form, so it seems more likely that the conversion was to a reverse implication.

This brings us back to one of the most important points made in the introduction to inference tasks: that people's reasoning on conditional rules is critically dependant on their interpretation of those rules. The point is strengthened by a subsidiary task in Evans' experiment in which subjects were required to construct thematic examples of 'If then' and 'only if' sentences: it was found that most of these sentences involved temporal or causal connections which were only 'natural' in the one form of the rule. These connections were such that when the antecedent event preceded the consequent event in time, the relationship was expressed in 'If then' form, but if the reverse was true, the 'only if' form was used. This line of enquiry was pursued further in another paradigm, the truth-table task, by Evans & Newstead (1977), in a study aimed not only at enlarging on Evans' findings but also at distinguishing conversion and interpretational explanations. It will be considered in more detail in the next chapter.

Of course, the interpretational effect which has drawn most attention is the possibility that subjects will make 'illicit conversions' of rules of implication into rules of equivalence. Wason & Johnson-Laird mentioned several factors which may lead to this, including negating the antecedent. The linguists Geis & Zwicky, on the other hand, turn the problem around: they contend that the interpretation of a conditional as an equivalence, which they call "conditional perfection", is a natural tendency in language, in their

words, "conditionals are understood to be perfected unless the hearer has reason to believe that the converse is false". (Geis & Zwicky, 1971). Thus, in another of their phrases, the conditional "invites the inference" that it implies its converse unless it is in a context that dictates otherwise. Geis & Zwicky come to similar conclusions as Wason & Johnson-Laird with regard to the classes of sentences which clearly do imply their converse, but are not so sure about the influence of syntax. On this argument it should not be too surprising that people tend to make the AC and DA inferences in abstract tasks which are precisely intended to be context-free. Two studies have looked specifically at these fallacies, which perhaps are not fallacies at all in natural usage. One reports a relatively uninteresting evaluation inference task using negative rule components: Roberge (1974) found that NA rules produced the most 'errors' - i.e. the most DA and AC inferences evaluated as true - and that more AC's than DA's were affirmed as true. Both these findings have been confirmed in other studies, detailed above. The other work, however, that of Wason (1964), is of much greater importance here because it mounted a two-pronged attack aimed at disambiguating the conditional. The first of these is the use of thematic materials, which most writers agree should have the desired effect; in this case, they did not seem to, as 33% of Wason's subjects initially succumbed to DA and 67% to AC, proportions which are well in line with studies using abstract materials, but which contrast sharply with the unpublished study by Shapiro, reported by Wason & Johnson-Laird, which used thematic materials and on which "hardly any errors were made" (p. 56). This aside, Wason's innovation was to use the inference problem embedded in a procedure which allowed several successive inferences to be made within the same task. One group of subjects could make

logically valid conclusions which were consistent with both previously made and succeeding fallacious conclusions, while another's valid conclusions conflicted with the previously made fallacies. Thus the second group of subjects were led to contradict themselves. This group showed a significant tendency to stop committing the fallacies, but the other group continued to make them, an illustration of how actually using the inferences served to disambiguate the conditional. It should be remembered, of course, that only half the subjects committed the fallacies in the first place. Perhaps a milder interpretation of Geis & Zwicky (1971) is called for: some people will interpret a conditional as an equivalence, but situational variables can cause such interpretations to be changed. Some studies by Taplin and Staudenmayer which look specifically at this question will be examined shortly; particular issues are involved in their case and it would be out of place to spend time on them here.

Disjunctives

At this point it will be useful to digress from conditionals for a while and consider disjunctive ('Either...or...') inferences, since similar logical and psychological issues are involved, and similar experiments have been done on them. These experiments are not so numerous as those on conditionals, and the bulk of them have been conducted by Roberge, sometimes in direct comparison to conditionals. Indeed the prime motivation for looking at disjunctives seems to have been to provide a different slant on the work with conditionals, as for every conditional sentence, there is a logically equivalent disjunctive. The relation between a conditional and a disjunctive is achieved through the application of negation, and a simple example will suffice. 'If p then q' is logically equivalent to 'Either not p or q',

since in both cases, given p we must conclude q, and given not-q conclude not-p. The fallacies also apply, if the disjunctive is assumed to express inclusion, i.e. if it is taken to mean'Either p or q, or both'. If it is taken to express exclusion, i.e. 'not both', then the disjunctive is logically interchangeable with an equivalence conditional, as all premises lead to a valid conclusion in both cases. The inferences for the unnegated disjunctive are shown in Table 2. However, it may seem that the substitute for the AA conditional, being an NA disjunctive, is harder to follow, and if this raises suspicions that the logical and psychological substitution might not marry, that is all to the good, as those suspicions are about to be confirmed.

Roberge (1974) compared DA and AC inferences on both conditionals and logically equivalent disjunctives. He found that while the NA sentence produced the most errors on conditionals, all the disjunctives containing negative components were more difficult than the AA disjunctive, and that most fallacious inferences on the disjunctive were made on negative rules when the argument involved affirmation of the first component-- logically equivalent to denying the antecedent of the AA and AN conditional rules. This is not surprising, since we have already seen how subjects find it more difficult to appreciate that an affirmative denies a negative, than that a negative denies an affirmative in conditional tasks. In a later experiment comparing inclusive and exclusive disjunction (Roberge, 1976a), using an evaluation task as in the 1974 experiment, and investigating the denial of the first component, which leads to a valid conclusion, Roberge repeated the results for inclusive disjunction, but found the order of difficulty for exclusives to be slightly

different: AA and NN rules were easier to reason with than AN and NA rules. Exclusives produced generally fewer errors, and the facilitating effect of explicit denial (of an affirmative by a negative) did not transfer to exclusives. It is more legitimate to talk about errors in these experiments because interpretation which subjects should apply to the disjunctive is not assumed but specified in the appendage - 'or both' or 'but not both' - to the sentence. In another study using systematically negated sentences (Roberge, 1976b), this time on exclusive disjunctives only, and using all four possible inference forms, Roberge confirmed and added to his previous results: single-negative rules were again more difficult, but this time AA was easier than NN; denial of a negative was most difficult to deal with, especially when the correct conclusion was an affirmative (cf. Evans 1972a, 1977a). In all these studies with abstract materials then, there are some fairly consistent overall findings which accord well with previous results from conditionals: negatives cause difficulty, especially if they have to be denied, though it is single negatives which cause the most difficulty. What of thematic materials?

The first experiment dealing with thematic materials was also one of the earliest reported studies of reasoning with disjunctives. Johnson-Laird & Tridgell (1972) used a construction task, with subjects supplying their own solutions to three disjunctive arguments; the disjunctives were not specified as to inclusive or exclusive interpretation. All arguments involved the denial of the second component, but the nature of the denial differed between them: the first was a direct denial of an affirmative rule component by a negative minor premise, the second an implicit denial by an antonym, and the third the denial of a negative component by an affirmative, called by the

experimenters an "inappropriate negative". On the basis of previous research with negatives they predicted, and found, that the first problem was easiest, both in terms of error rates and solution times, the third the most difficult, and the second intermediate. Their explanation was that it is easier to grasp that a straight negative denies an affirmative, slightly more difficult to use the implicit negative because of a presumed extra step to convert it to a straight negative, and much more difficult to grasp that an affirmative denies a negative. However, these conclusions are not entirely satisfactory bearing in mind Roberge's findings that a negative anywhere in a disjunctive seems to create particular difficulties, probably, as Evans (1972c) argues, of interpretation. Roberge (1977, 1978) has mounted a detailed investigation of the interaction of different types of materials with other factors known to influence disjunctive inferences: negation, inclusive/exclusive interpretations, and affirmation and denial of the first and second components. These studies used both abstract and thematic materials (capital letters and sentences such as 'Either Joan is intelligent or she is rich, or both'); the 1977 study also used 'contradictory' thematic sentences (e.g. 'Either John is intelligent or he is stupid, or both'). It seems that the latter were treated, not surprisingly, as exclusive rules in spite of the 'or both'. The 1977 study found no difference between materials: in both cases there were fewer logical errors on the valid denial of the first component argument compared with the fallacious affirmation of the first component argument. The 1978 study, using denial of the second component only, and including negated rules and exclusives, again found no overall effect of materials, although there was an interaction between materials and polarity: NA

sentences were easier with thematic materials, NN with abstracts. Once again a single negative caused particular difficulty (less so with thematic materials), and explicit denial led to fewest errors. There were more 'indeterminate' responses with inclusive sentences than with exclusives, and of course there should be none in either, since denial of the second component leads to a valid conclusion under both interpretations, so this result seems to indicate that inclusive disjunction makes a less definite statement than does exclusive disjunction. Overall, this study confirms Roberge's earlier findings on inferences involving the first component of the rules and extends them to inferences on the second component.

Before leaving disjunctives, some comment should be made, in view of the findings reported above, on the use of conditionals and disjunctives as each others' equivalents in reasoning experiments. This usage is justified on one count but unjustified on three others, and the split coincides nicely with that between the logic and psychology of reasoning. The justification rests on a fact of logic, viz. that 'If p then q' and 'Either not p or q' are logical correspondents: they both express material implication or equivalence, depending on their specification (logicians would assume implication, as they also assume the inclusive interpretation of the disjunctive). Hence they both entail the same rules of inference and are falsifiable by the same truth-table case(s). However, the three other counts, which are empirically based, promote the conclusion that the correspondence between the two rules has no psychological reality. In the first place, there seem to be fundemental differences in interpretation between the two rules: it has been found that any negation causes difficulty with disjunctives, whereas it is only

negation of the antecedent which does so with conditionals. It should also be remembered that the MP conditional inference holds up across all rules, with or without negation; none of the disjunctive inferences does so. Secondly, on a related point, it is clear from the research on negation that there are bound to be differences in the use of conditional and disjunctive rules, since on the implication conditional one of the valid inferences arises from an affirmation and the other from a denial, while on the inclusive disjunctive both valid inferences arise from denials. Thirdly, there is the well-established directionality of the conditional, which we have seen to be an important factor in conditional reasoning. This does not apply to the disjunctive, inclusive or exclusive, since these are rules of alternation, and do not proceed from the establishment of an antecedent to the conclusion of a consequent. If there is any interchanging, it should be between equivalences and exclusives, since directionality plays less of a part in the former, and the inference rules (and truth-table cases, as we shall see) are symmetrical.

This issue focuses on the disparity between logic and psychology, and particularly on the dangers of assuming the psychological reality of logical constructs. Attempt have been made, notably by Piaget and most recently by Braine, to derive systems wherein logic and psychology combine; this approach is dealt with in detail in the Discussion. Even when not used to promote a theory, an identity of logic and psychology has been assumed in some studies, and the offenders in mind here are those of Taplin, Staudenmayer, Rips, and Marcus. These were inference studies, but they were used to infer truth-tables. These studies form the opening to the next chapter; the next chapter is about truth-tables.

## CHAPTER 3



Tables

In the previous chapter on inference tasks, the processes under investigation were those involved in deciding which conclusions could or could not validly follow from the establishment of one or other of the components of a conditional or disjunctive statement, or their negations. For the purposes of this deduction questions as to the truth of the statement are set aside - it is irrelevant to the logical validity of the arguments whether or not the statement is a true or false one in relation to fact. In this chapter, we consider a set of tasks relating to the logical process of deciding whether or not a statement holds true. The concern is not with factual truth, but with the truth or falsity of the statement given the occurrence of instances relating to its components. Each conditional or disjunctive rule has two components which can either be affirmed or denied; they might both be true, one true and the other false, or both false. There is thus a logical truth-table of four instances. The effect of each instance on the logical truth status of the rule is related to the extent to which each allows the valid inferences, and we can return to our tigers again for an example of how this relation applies for a rule of implication. We know by the MP inference that if there is a tiger there must also be stripes, and from the invalid AC and DA inferences we know that there may be stripes whether or not there is a tiger present; the rule does not legislate on the actual presence or absence of tigers, only what is conditional on that presence. However, we also know by MT that if there are no stripes, there can be no tiger, and it is therefore a straightforward deduction that the only combination of characteristics which could decisively falsify the rule itself would be a tiger without stripes. All other combinations - striped tigers or non-tigers with or without stripes -

allow the rule to stand as true. A rule of equivalence behaves slightly differently, as it implies its converse; consider the rule 'If it is an elephant, then it has a trunk'. Of course, it would not be an elephant if it did not have a trunk, but equally if it does have a trunk, it cannot be anything,but an elephant (American cars and large packing cases being excluded from this definition of 'trunk'). Therefore not only elephants without trunks but also trunks on things other than elephants falsify this rule. The two functions are summarised in Table 5.

It is the relation of this logical truth system to the ways in which people actually go about evaluating the truth of these rules which is the subject of research using truth-table tasks. There are studies which look directly at this process by getting subjects to construct or evaluate the various instances, but first we need to examine a set of American studies which aimed to integrage inference and truth-tables psychologically, an approach which, it will be contended, was mistaken. These studies used a common methodology, of observing performance on inference tasks and from that inferring subjects' truth-functional interpretations of conditional sentences.

Inferences and truth-tables do not mix

The methodology was established by Taplin (1971). He used the familiar paradigm of the evaluation inference task - a complete chain of argument which the subject must evaluate as valid or invalid. In this case, the subjects were to judge whether the conclusion necessarily followed from the premises and answer yes or no accordingly; there was no 'indeterminate' option. Taplin used thematic materials and also a logical falsehood in an MP argument to check for any bias towards affirmative responses. No negative rule components were used,

TABLE 5 The truth function of the conditional. Formal and 'defective' truth-tables for implication and equivalence are shown

| | | Truth status of 'If p then q' | | | |
|---|---|---|---|---|---|
| | | Formal | | 'Defective' | |
| Given: | Truthtable case | Implication | Equivalence | Implication | Equivalence |
| p and q | (TT) | True | True | True | True |
| p and $\overline{q}$ | (TF) | False | False | False | False |
| $\overline{p}$ and q | (FT) | True | False | Irrelevant | False |
| $\overline{p}$ and $\overline{q}$ | (FF) | True | True | Irrelevant | Irrelevant |

Letters in brackets express truth values for antecedent and consequent items; T = true, F = False.

$\overline{p}$ and $\overline{q}$ are the negations of p and q.

and twelve problems of each type of argument were presented, giving each subject 60 deductions to make. The frequencies with which each inference was evaluated as true are not reported, but it is possible to extract them by laborious computation from a distribution table; they are: MP: 92%, DA: 61%, AC: 57%, and MT: 63%. Except for DA, which is rather high, these results are consistent with previous findings (see Table 3). The logically false MP argument was rejected 94% of the time, showing a lack of 'affirmation bias'. Taplin does not report these frequencies because he is not concerned with them; his analysis is based on the assumption that "if it is known that a given conditional sentence is true, and the truth value of either the antecedent or the consequent is also known, then the truth value for the given conditional sentence may be derived from judgements regarding the validity of a conclusion involving the consequent or the antecedent respectively". (Taplin, 1971). The key word here is "judgements", which is an expression of a (presumed) psychological process; if the sentence simply said "derived from the validity..." etc., the argument would make perfect logical sense, as in the illustration with tigers above. The ensuing discussion attempts to show that, once again, this confusion of logic and psychology has led to faulty interpretations and, ultimately, faulty theorising. But first to the analysis.

The frequencies with which the inferences were made are given above, but Taplin uses a logical derivation of truth function from inference, coupled with correlations between responses on certain inference pairs, to deduce individual truth functions for each subject. For an index of logical performance he takes consistency of responding, i.e. deviation from chance frequencies on each problem; noting a lack of uniformity here, he concludes that there is no truth function for

conditionals which all individuals use. However, there were significant correlations of performance between some inference pairs, and he proceeds from this to an individual analysis of subjects' consistency of responding. Only 48% of subjects were consistent on all four inferences, most (37.5%) of these consistently affirming all four as true; only 3.6% correctly (for implication) dismissed DA and AC while assenting to MP and MT. It was therefore concluded that 37.5% of subjects were using the equivalence truth-table and 3.6% the implication truth-table (see Table 5), leaving 58.9% without a formal truth-table. The large number of inconsistent subjects - over half - led to a second investigation controlling for the possible role of conclusion plausibility and sentence length: Taplin & Staudenmayer (1973) repeated the experiment using abstract materials (letters). They also included the other three logically opposite conclusions, beside the MP falsehood, so there were now eight forms of argument for the subjects to evaluate. Using the same analysis, 20.8% of subjects were found to be inconsistent on at least one of the arguments, and of the consistent ones the great majority were again ascribed an equivalence interpretation. This is reflected in the differences in inference frequencies (again derived and not presented). For the normal arguments, i.e. excluding the opposites, they were: MP: 99%, DA: 82%, AC: 84%, and MT: 87%. This quite startling difference from the Taplin (1971) results is attributed to the use of abstract materials, but a glance at Table 3 will also show that these results are some way out of line with others on abstract rules. However, the results are interpreted as supporting Taplin's, as the majority of inferred truth functions are again of equivalence. Recognising a shortcoming in the two experiments - they had compounded 'logically false' with 'indeterminate' both of

which should have been contained in the 'no' response, and that this makes the experimenters' truth-table inferences questionable - a second experiment was conducted including an 'indeterminate' response option and a corresponding adjustment in the truth-table inference structure to accommodate it. There is no mention at all of inference frequencies in the results, but on their analysis there were more statistically consistent subjects (37.6%), though not as many as in the Taplin experiment, and a dramatic change in the proportions of the truth-tables: this time 33.8% of subjects came under the implication category and only 13.7% could be assigned to equivalence. This difference is explained as being due not only to the experimenters' possible misinterpretation of the binary response categories, but also to the possibility of the third category implying, by its presence, that it should be used somewhere.

Staudenmayer (1975) extended this line of work in an effort to elucidate the role of such factors as these affecting subjects' inferred interpretations of conditional sentences. The four factors used in his investigation were materials (abstract or thematic), form of the connnective ('if...then' or 'cause'), semantic relation (anomalous or causal), and the relation of the antecedent to the consequent (necessary or not necessary). The task and analysis were based on the second experiment of Taplin & Staudenmayer (1973), and again inference frequencies are nowhere reported. The finding of higher consistency with abstract than thematic rules was repeated, although on inspecting the data table this seems to be largely due to a very high degree of consistency of responding on abstract-causal sentences rather than to an overall tendency. When only the 'ordinary' thematic and abstract rules are compared - the ones

corresponding to those to which the terms have been applied in the literature - consistency is almost identical between them. Cause and necessity sentences seemed to produce more equivalence classifications. On 'if...then' sentences there was little difference between equivalence and implication ascriptions due to abstract or thematic content, except in the 'necessity' condition, where equivalence predominated. Thus the finding of Taplin & Staudenmayer (1973) that there were more implication interpretations on 'if...then' conditionals with abstract content was not confirmed. These results do not seem totally conclusive. Such high and variable degrees of consistency of responding cast doubt over what are in any case equivocal findings; it seems possible to conclude that there is, as these authors say, no evidence of any systematic relationship between the logic of the conditional and its interpretation by subjects in reasoning tasks. However, the main criticism of these experiments rests not with the findings, but with how those findings were arrived at. Staudenmayer (1975) presents a highly revealing decision tree diagram which illustrates Taplin's principles of deriving truth-tables from inferences. It is revealing because it rests entirely on logical principles. Does this logical correspondence therefore have psychological validity? These researchers seem to be saying that truth-tables thus derived reflect the subjects' interpretations of the rules on which inferential operations are then based. If this is the case, should not formal logic as a whole provide an adequate explanation of deductive reasoning? All the research reviewed in this and the preceding sections must cast doubt on this idea. Surely the only way to see whether people's (assumed) interpretations of a rule in an inference task correspond to their interpretations of that rule in a truth-table task is to mount a direct comparison of those tasks.

Rips & Marcus (1977, Experiment IV) did just that. They first gave their subjects a truth-table task - evaluating the truth status of a rule against its possible paired truth-table instances, as set out in Table 5 - followed by an evaluation inference task, performing a Taplin-style analysis on the latter. Even though they only include in this analysis data from subjects classifiable under implication or equivalence truth-tables, they found that only just over half of these subjects could be so classified on the basis of their inferences. Versions of the task with the three-way and two-way responses (i.e. with and without an 'indeterminate' category) were conducted, and these are the inference frequencies:

| | % | | | |
|---|---|---|---|---|
| | MP | DA | AC | MT |
| 2-way | 100 | 21 | 23 | 57 |
| 3-way | 99 | 31 | 29 | 62 |

Rips & Marcus construct an elaborate 3-stage model with several "error assumptions" to account for this disparity, but the most obvious explanation is that the tasks are measuring different things. Taplin himself alludes to this in his 1971 paper when, noting a discrepancy between his results and those from other paradigms directly concerned with the truth function, he recognises that an experiment in which subjects are told to assume that the rule is true, such as his and other inference tasks, may be psychologically distinct from one in which subjects are asked to test whether the rule is true or not. As there is good evidence from another paradigm that subjects may do inconsistent things on the same task (Wason & Johnson-Laird, 1970; Wason & Golding, 1974; Wason & Evans, 1975), it would not be at all surprising if they were to do different things on different tasks. Rips & Marcus have demonstrated that indeed they do. Perhaps the most interesting question here is how such widely differing data, as

revealed in the inference frequencies before they are transformed into truth-tables, come to be invoked as evidence for the same theoretical position. The moral of this particular story must be that one should use inference tasks to investigate inferences and truth-table tasks to investigate truth-tables. We have seen what happens on inference tasks, and now we turn our attention to truth-table tasks proper.

The 'real' task

Truth-table tasks are concerned almost exclusively with conditional sentences. Their increasing use over recent years stems partly from the general disaffection of psychologists with the formal calculus as a model of thought, and more particularly from a pilot study briefly reported by Wason in 1966, in the debut of the Selection task. Wason reported that subjects seemed to be using a third category in evaluating conditional sentences - irrelevance. Formal logic, of course, specifies a bivalent truth function, where an instance either verifies or falsifies. The instances in which the antecedent component was falsified tended to be disregarded as irrelevant to the rule; for instance, in our example of tigers and stripes, lions with or without stripes would be regarded as having nothing to do with the rule - not verifying it, as implication would dictate. Wason referred to this kind of truth-table as 'defective' (see Table 5). The suggestion that people treat rules of material implication in a non-truth-functional manner led to a number of studies as to why this should be. Wason (1968), in the first published paper on the Selection task, reports additional tests in which he attempted a direct derivation of the truth-tables underlying the selection task. These tests were designed as 'therapies' to facilitate performance by explicating the structure of the task to the subjects, who would then proceed to the task with this new-found insight and get it right

(in fact they did no such thing, but more of that later). One was the 'projection' of falsifying values on to non-selected items after a first attempt at the Selection task, but another involved evaluation of cards bearing each of the four possible antecedent-consequent combinations as they affected the rule, a much more 'direct' method. Some confirmation of the 'defective' truth-table was found, although many subjects seemed to regard only the doubly negated item (FF) as irrelevant, with the false antecedent/true consequent pair (FT) falsifying. This seems to constitute evidence for defective equivalence as much as implication. The interpretation of these results is risky though, since 'hints' were given to the subjects, and it is hard to be confident about deriving truth-tables from this study. The most direct method of testing this is to present the rule and ask the subjects either to compose instances which verify or falsify it, or to ask them to evaluate the rule in the face of the four possible truth-table cases, as set out in Table 5.

This latter procedure was used in an experiment by Johnson-Laird & Tagart (1969) with two objects in mind: to see whether Wason's observation of the use of an 'irrelevant' category would be confirmed, and to assess the extent to which the mode of expression of material implication affected its interpretation. Rules were abstract, concerning letter-number combinations on cards, e.g. 'If there is an A on the left, then there is a 7 on the right', and the instances were cards with a letter on the left and a number on the right, or with blank spaces or geometrical shapes as alternative falsifying items. The subjects' task was to sort each card into piles corresponding to the 'true', 'false', and 'irrelevant' cateogories. It was found that for the 'If A then 7' rule the most common (79%) classification

was for the card bearing A and 7 (the true/true, or TT, case) to be classified as true, the card bearing A and something other than 7 (true/false, or TF) as false, and not-A with 7 (FT) or not-7 (FF) as irrelevant - a strong confirmation of Wason's (1966) observations. This classification was also in the majority (58%) when the statement was expressed as 'There is never an A on the left without there being a 7 on the right', but the other two alternative sentences - 'There isn't an A on the left, if there isn't a 7 on the right' and 'Either there isn't an A on the left or there is a 7 on the right' - yielded no systematic pattern. The experimenters conclude that "the way in which implication is expressed exerts a decisive influence upon what it is understood to denote". This may well be so, but there are grounds for doubt as to whether the procedure used provided a valid test: we have already encountered the problems people have in interpreting the singly-negated disjunctive, and the lack of consistency of responding to it here seems to confirm this. Similarly, there is a varying use of negation between the alternative rule forms, which may in itself have contributed to the performance differences on them. Logically, they may all express implication, but of course psychologically their meanings may vary or even, in the cases of 'Not-p or q' and 'Not-p if not-q', dissipate completely. Secondly, it could be argued that the 'irrelevant' responses were cued by the presence of the third category - most subjects would not think it was there to be ignored. However, the lack of use of this category with the disjunctive rule form argues against this idea, as do the findings of an experiment by Evans (1972b) which sought to control for both this and the negation factor.

Negatives

This experiment used the expedient of systematically negated conditional rule components, which first appeared in Roberge's (1971a, b) study of inferences. The subjects' task was to construct instances which could verify or falsify a given rule, selecting items from an array of prepared cards, and the ingenious aspect of this procedure was that it was exhaustive: each subject was asked to compose instances until he judged there were no more. In this way, any unused combinations could be inferred to have been irrelevant - the third category was not cued. By applying systematically negated components the roles of truth value and negation could be separated. This neatly balanced design is central to much of the ensuing review and discussion, so a digression to explain it fully is appropriate here.

In the ordinary double-affirmative (AA) 'If p then q' rule, truth/falsity and affirmation/negation are compounded, i.e. a true value is an affirmation of an item, a false value a negation. The separation of these characteristics can be illustrated by looking at the true antecedent/false consequent (TF) case, which is logically (for implication) the only falsifying instance. In the case of the 'If A then 7' rule (see above) this is represented by A and, say, 8, or p and $\bar{q}$ (not-q), and so the true/false pair is also the affirmative/negative pair. Consider however a rule with a negative in it: 'If there is an A then there is not a 7' (an AN rule). Here the TF case would be A and 7, or p and q, a double-affirmative, or double-matching, pair, so truth and polarity are separated, at least on the consequent. For an 'If not A then 7' (NA) rule, the TF case becomes, say, B and 8, or $\bar{p}$ and $\bar{q}$, and for an 'If not A then not 7' (NN) rule the TF case is B and 7 ($\bar{p}$ and q). Thus each truth-table case is represented

on the four rules by a different combination of items, and of course each combination therefore has a different logical value for each rule. This is summarised in Table 6, where the four possible affirmed and negated item combinations are given their appropriate truth-table value for each of the four rules. The notation used in Tables 5 and 6 will be adopted in the ensuing pages for brevity, and to avoid confusion item combinations will be referred to by their logical value (TT, TF, etc.) rather than their matching status (pq, $p\bar{q}$, etc.) as items in a rule, unless a specific point is being made on this status.

Matching bias

In Evans' (1972b) experiment, each subject was given all four rules and had to construct his own truth-table cases; the rules were abstract and concerned coloured shapes. Two important trends in the data emerged: firstly, the TT case was almost unanimously constructed as a verifying instance on all rules, and the great majority of subjects also constructed the TF case as falsifying. Secondly the FT and FF cases were usually left out as irrelevant on the AA and AN rules, but were usually constructed on the NA and NN rules. Examining the incidence of 'irrelevant' items more closely, Evans found that these tended to correspond to items which did not match the values named in the rules, while matching items tended to be constructed, e.g. the FT case to falsify the NA rule - a double-matching, pq, item. These findings are important: they confirm the existence of a response bias, here termed 'matching bias' by Evans, concurrent with a logical tendency, for after all, most subjects correctly constructed the TT and TF cases. In a later experiment, Evans (1975) repeated this procedure with an evaluation

TABLE 6 The relation between truth-table case and named (matching) items in the four conditional rules with systematically negated components. Notation as in Table 3.

| | | Truth-table case | | | |
|---|---|---|---|---|---|
| | Rule | TT | TF | FT | FF |
| AA | If p then q | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ |
| AN | If p then not q | $p\bar{q}$ | pq | $\bar{p}\bar{q}$ | $\bar{p}q$ |
| NA | If not p then q | $\bar{p}q$ | $\bar{p}\bar{q}$ | pq | $p\bar{q}$ |
| NN | If not p then not q | $\bar{p}\bar{q}$ | $\bar{p}q$ | $p\bar{q}$ | pq |

task, where subjects had to classify the four instances as true, false, or irrelevant for each rule. Almost identical results were obtained, showing that the response tendencies observed here were not a product of the construction task itself. These results are in line with what we have already seen on quantified syllogisms and inferences, and they provide further evidence as to the divergence between logic and cognition on these tasks: not only do subjects depart from the truth function in their inclusion of an 'irrelevant' category, they are also influenced by non-logical task variables.

These ideas were elaborated by Evans (1972c) in a theoretical paper which criticised the approach of psychologists adopting formal logic as a competence model. The idea of using formal systems to gauge 'correctness' in reasoning tasks has already been referred to in the previous chapter: the matching bias results of Evans (1972b, 1975) confirm and add to the disquiet, establishing as they do the existence of strong non-logical determinants of performance on a straightforward logical task. Matching bias in the truth-table task is an embarrassment to those writers (e.g. Legrenzi, 1970; Rips & Marcus, 1977) who propose that the task gives a 'true' measure of people's truth-functional interpretations of conditionals and use it to explain the interpretations 'underlying' performance in other paradigms. Evans (1972c) suggests that formal competence models (which here would also include 'natural' systems such as that formulated by Braine, 1978) should be rejected in favour of a non-logical, two-factor account of reasoning: that performance is determined by the interaction of <u>interpretative</u> and <u>operational</u> factors inherent in a task. Interpretative factors have to do with the subjects' understanding of the premises of a deductive argument, and operational

factors pertain to the reasoning processes carried out. Examples of both can be drawn from Evans' matching bias studies: the overall tendencies to classify the TT and TF cases as verifying and falsifying would seem to reflect subjects' interpretations of the rules, but the influence of matching bias shows an operational effect, as it cuts across the logical consequences of certain truth-table cases. However, inferring the play of these factors in this way invites circularity, and Evans, mindful of the danger, argues that the two factors can only be distinguished by looking at different situations where one or other variable is controlled for. Thus if one factor is held constant, one can justifiably implicate the other in any effects observed in the data.

Using this approach, Evans compares results from inference experiments by Johnson-Laird & Tridgell (1972) and himself (1972a) which both used denial of the second component of a deductive argument, i.e. the same operation, but found different results, which are therefore taken to indicate an interpretative difference between the two experiments' materials. He uses the corresponding mode of attack to affirm the empirical genrality of matching bias in a later experiment (1975), not only by using an evaluation task rather than a construction task (see above), but also by using the logically equivalent 'p only if q' rule form. Once again, significant matching bias tendencies were observed, showing the influence of an operational variable, but there were detail differences between the 'if then' and 'only if' forms, especially in the rules with negated antecedents. These differences are presumably due to the interpretational differences noted in Evans' (1977a) inference task study between the two forms. However, the materials being abstract and therefore ambiguous, it

was necessary to look further at this question since, as Evans (1977a) had observed, both may express material implication but the 'only if' form is used most readily to emphasise the necessity of the consequent, while 'if then' seems to emphasise the sufficiency of the antecedent. Perhaps this was the interpretational difference responsible for the interpretative effect in Evans' 1975 study. The way to test this is to make the difference explicit. One could do this by using thematic sentences expressing the temporal/causal relationships outlined above and inherent in subjects' constructions of 'if then' and 'only if' sentences, or one could 'tag' abstract sentences to the same effect. Evans & Newstead (1977) did the latter: they retained abstract materials (letters) but used them in sentences expressing alternative temporal relationships, one where the antecedent event preceded the consequent event, and the other stating the reverse order. It was predicted that the first order would be more naturally expressed by an 'if then' conditional and the second by an 'only if', and that sentences with the time order matching the rule structure should therefore be easier to understand than sentences expressing an 'inappropriate' temporal relation. This interaction between rule form and temporal order was investigated by using a procedural innovation adapted from a design by Trabasso, Rollins & Shaughnessy (1971): split response times. The rules and instances were presented in separate tachistoscope fields, and the subject controlled their presentation himself. First he pressed a button to view the rule, then he pressed again for the instance, and finally he pressed a response switch to record his decision, true, false, or irrelevant. The rule-instance and instance-decision latencies were timed, and in this way the time taken by the subject to understand the rule (comprehension time) could be separated

from the time it took him to evaluate it against the instance (verification time). Both measures were found to reflect the predicted interaction. Of course, it is possible that subjects were simply converting 'only if' rules into converse 'if then' rules, and that all the observed interaction showed was that the antecedent-before-consequent order was easier to process. A significant overall tendency to convert 'only if' rules was found, in the frequencies, but the effect seemed to be, limited to those rules with negated antecedents. The results of this study do not add a great deal to what was already known about conditional truth-tables: the role of matching bias was confirmed, negated rule components having been used again, and the suggested interpretational difference between 'if then' and 'only if' conditionals, found before on inferences, was confirmed experimentally. The particular interest of this experiment lies in the ingenious procedure of splitting comprehension and verification times, since this allows the separation of interpretative (comprehension) and operational (verification) variables.

Content

The studies reviewed above have put rather a lot of weight on the role of non-logical response biases; some other work has concentrated more on interpretation, and it is to this which we turn now. Interpretative effects could be demonstrated by varying the materials in an experiment in which the operational effects are already known, any difference being most likely due to interpretation. There is some evidence along these lines: Legrenzi ( 1970) suggested that a conditional problem in a strictly binary or causal context, where the antecedent and consequent items come from populations of exactly two or when the antecedent is seen as the cause of the

consequent, would lead to an equivalence interpretation of the rule, and that this might be disrupted by a linguistic formulation of implication such as 'Not p and not-q', which does not share the temporal/causal connotations of the 'if then' form. This was investigated in an elegant experiment involving a pinball-type apparatus and rules about the passage of the ball and its consequences. The rules were (e.g.) 'If the ball rolls to the left, then the green lamp is lit' or 'It is not possible for ball to roll to the left and the green lamp not to light', which are both expressions of 'left implies green'. There were just two channels, left and right, and two lights, red and green. The four truth-table cases were presented to the subjects, e.g. a ball rolling to the left and the red light coming on (TF), who classified them as compatible with the rule, incompatible, or irrelevant. Sure enough, both predictions were confirmed: 75% of subjects classified the cases according to the truth-table for material equivalence (see Table 3) under the 'if then' rule, only 17% judging FT and FF irrelevant, while under the 'Not p and not-q' rule only 10% adopted equivalence; 27% classified according to the defective implication truth-table, and 63% were consistent with material implication. These results are taken by Legrenzi to show that it is the causal connection and not the binary situation that promotes equivalence interpretations.

Rips & Marcus (1977), in their wide-ranging article, continue this line of research. Noting the differences between the results of Legrenzi and Johnson-Laird & Tagart (1969), they propose that there could be several reasons: not just the binary-causal task used by Legrenzi, but also the materials, or language and population variables (Legrenzi's experiment was conducted in Italy). They therefore

replicated both experiments and added their own materials condition, a thematic one of rules about the colours and markings of tropical fish. (This third condition is most like a thematic condition in the sense adopted here. Legrenzi's materials are not totally abstract, but neither are they thematic in the sense of being realistic rules about everyday situations). They also made the populations of antecedent and consequent binary or non-binary by restricting them to two or three items. So far so good, but unfortunately the effect is spoiled by the exclusion of an 'irrelevant' category and a peculiar analysis in which equivalence classifications are lumped together with those consistent with an interpretation of the rule as a conjunction ('p and q', where TT is true and all other cases false). They found the highest number of 'equivalence' classifications on the Legrenzi materials, but no difference due to the binary/non-binary factor, so like Legrenzi they come down in favour of the causal connection as being behind equivalence interpretations. In this experiment, 88% of subjects were classifiable under implication or 'equivalence', but as the authors concede, this high figure is most likely due to the absence of the'irrelevant' category.

In a further experiment they explored the possibility that it is an inferred correlation between antecedent and consequent rather than an explicitly causal connection which is promoting the equivalence interpretations. Using the same paradigm, they specified in the instructions in a 'correlated' condition that the antecedent was associated with just one consequent value, while in an 'uncorrelated' condition the instructions stated that the antecedent might go with any of the consequent values. There was no difference found in the number of 'equivalence' classifications between the materials this

time, which might have worried the experimenters but does not seem to, but a significant increase in them in the correlated condition across all materials. Only the Legrenzi materials could be construed as being in any way causal, so from these results it looks as if it is an inferred correlation between the antecedent and consequent of a conditional which encourages equivalence interpretations.

Rips & Marcus' paper has a large theoretical content, and they interpret their results in terms of their Suppositional theory of conditionals. This takes its cue from Wason's analysis of negation; it will be recalled that, according to Wason, the natural function of a negative is to deny a presupposition. Rips & Marcus take a similar line in proposing that a conditional is interpreted in terms of its inherent suppositions, a supposition being a sum of "the current data base and a single 'seed' proposition". The current data base is the universe of (relevant) things we consider to be true, the 'seed' the hypothesis contained in the antecedent of the conditional (cf. Wason & Johnson-Laird, 1972, p. 90: "The antecedent is an explicit statement of a presupposition"). Compared with formal logic, the suppositional idea works quite well, as it can account for the role of prior beliefs and the defective truth-table. The theory rests both on these and on the "asymmetry" (directionality) of the conditional, evidence for which has been well documented and which they enlarge. Thus the results above are taken to show that "the crucial factor...is the form of the relation believed to obtain between Antecedent and Consequent values; in other words the function mapping the Antecedent onto the Consequent range" (Rips & Marcus, 1978). There is, however, a whiff of the post hoc about these ideas: it is not clear how the 'data base' is to be specified a priori, unless artificially as they do in their experiments, and they apply

no specific predictions which could not have been derived from other approaches. The biggest problem here is to devise an independent test of what is considered relevant and irrelevant to a given conditional.

The Suppositional theory does have the advantage of emphasising the role of content in testing the truth of conditionals, something advocated earlier by Wason & Johnson-Laird (1972). The predictive strength of both approaches is questionable, however, the former for the reasons stated and the latter because of a certain vagueness: beyond listing some pursuasive examples, they go no further than to assert the weakness of the formal calculus, the bewidering artificiality of abstract materials, and the neglect of presuppositions in conditional arguments. Evans' two-factor, multi-paradigm approach seems more promising, but beyond the suggestion that realism should strengthen the interpretative factor, it is not clear what should happen in judging the truth of realistic rules. The only truth-table experiments to use thematic materials are those of Rips & Marcus, and they do not use negated rules or an 'irrelevant' category, so the scope for testing for response biases is limited. Would matching bias influence the truth-table task if thematic materials formed the content? This question is one of the prime concerns of the research to be reported later, and fuller consideration of it will be granted at that stage.

In this chapter we have seen once again how subjects' performance on a logically structured task diverges from the performance required by logic. They do different things on tasks with the same underlying structure - thereby rendering unsound those studies which ignore this behaviour; they use a truth category alien to the formal system; and their responding is influenced by non-logical response

factors. It also seems that the content of the problems may influence truth-table task performance. However, this question has hardly been touched outside the Rips & Marcus investigations, and so evidence for Evans' contention about the role of content must, at present, come from another paradigm. Such evidence has indeed been provided, and in the next chapter its source is examined. The paradigm in question is Wason's Selection task.

## CHAPTER 4



Tables

The Selection task, devised by Wason and first presented in part of a general psychology book chapter in 1966, has succeeded in one respect in which the two preceding paradigms have not: generation of a large amount of research and theory in a short time. All the publications about to be reviewed have appeared in the space of 11 years. Its appeal for the experimental psychologist lies, as Wason has compactly noted, in the enigma of its structural simplicity and psychological complexity. Stripped to its components it stands revealed as indeed a very simple problem, but presented in its entirety to a naive subject it acquires a daunting complexity, leaving error and irrationality wherever it goes. As to a lesser extent do inference and truth-table tasks too, of course. All three have concentrated mostly on problems incorporating the logic of material implication; is the Selection task then an inference problem or a truth-table problem? It appears to contain elements of both, as can be seen from an outline of its basics.

Problem structure

In its prototypical form, the Selection task consists of a rule of material implication such as 'if p then q' which is given to the subjects along with four cards, each of which bears a different combination of p and q or their negations ($\overline{p}$ and $\overline{q}$), one value on each side of the card. The subject is allowed to see only one side of the card though, and his task is to select all and only those cards which he must fully examine to test whether the rule is true or false. The four cards show the values p, $\overline{p}$, q, and $\overline{q}$. The correct selections are the p and $\overline{q}$ cards; this is because the only decisive test of a rule is to see whether it could be wrong, not establish that it may be right. As we saw in the previous chapter, the only combination of items which could falsify a rule of implication

is p and $\bar{q}$, and the only cards which potentially carry this combination are p, which might have $\bar{q}$ on the other side, and $\bar{q}$, which might have p on the other side.

At first blush this looks like a truth-table task - subjects are, after all, asked to test the truth value of a rule in the light of certain instances, and indeed many early papers use a truth-table task as part of the procedure. However, it also resembles an inference task, by dint of its requirement of subjects to reason from one item to another, and is treated as such in Wason's first full publication on the problem (1968): the two correct cards are correct because they are the only ones from which a valid inference could proceed (MP and MT from p and $\bar{q}$ respectively), and differences in selection are regarded as differences in tendencies to make or withhold these inferences. Logically, of course, there is no problem, since the truth function can be derived from the inference rules, and vice-versa, and Wason (1977a) defines the task in this vein as "a 'meta-inference' problem - it requires a deductive inference about the conditions from which a valid inference could be made". This is one way, a logician's way, of looking at the task - a neat encapsulation of the derivation of truth function from inference rules, and anyone familiar with material implication would appreciate it. Perhaps, though, it is just this neatness which should activate caution in the psychologist: subjects in experiments are not so qualified, and we should be wary of presupposing one thing while perhaps asking the subjects another. We have already seen what difficulties can arise by assuming parallel logical and psychological correspondences between inference and truth-table tasks, and in the next few pages we shall see this divergence again, as the Selection task defies the

predictions of other paradigms and confirms that different tasks ask different psychological questions.

Experiments: abstract materials and therapies

To the observed behaviour then. In his original article Wason (1966) found that the correct response, p with $\overline{q}$, was rare. Subjects selected the p card readily enough, but hardly any selected $\overline{q}$, most selecting just p or p and q. This pattern remained even when the cards were fully exposed and the subject was asked which would prove the statement to be a lie (i.e. a truth-table task with the same materials). Over repeated trials the incidence of $\overline{q}$ selection increased, but the tendency to select q did not diminish. Wason attributes these errors to subjects regarding negated values as irrelevant (the first assertion of his 'defective' truth-table; see Chapter 3) and seeking only to verify the rule. His short account anticipates several lines of research: the use of truth-table tasks and other devices in an attempt to improve performance, and underlying this the formalistic assumption that subjects would appreciate the logic of the problem and get it right if only certain unknown obstacles were removed. Wason (1966) also anticipates a great deal of future data.

Most of the experiments and all of the theories have been concerned with the Selection task in its 'standard' form, outlined structurally above, and involving simple 'if then' rules with abstract materials. Below are summarised the selections of 369 subjects reported in 10 different papers in which this or a similar format is used (Wason, 1968, 1969a; Wason & Johnson-Laird, 1970; Goodwin & Wason, 1972; Wason & Golding, 1974; Johnson-Laird, Legrenzi & Legrenzi, 1972; Evans & Lynch, 1973; Bracewell & Hidi, 1974;

Gilhooly & Falconer, 1974; van Duyne, 1974). Experiments in which the data are presented in an unusable form, such as correct v. incorrect, are excluded. These are the percentage proportions of subjects who initially selected the common combinations of cards:

| | p | pq | $pq\bar{q}$ | $p\bar{q}$ | Others |
|---|---|---|---|---|---|
| % | 25.7 | 39.0 | 8.9 | 9.2 | 17.1 |

We can see that these patterns are very much as Wason (1966) appears to have found them, and this is remarkable considering what has been done to try and alter them (although it should be remembered that these figures do not reflect any procedural manipulations operating to change selections). Let us examine some of these manipulations before passing on to the theories which have attempted to account for the data.

A flavour of this approach can be got by considering the original paper by Wason (1968) which introduced the notion of "therapies" (his term) designed to facilitate logical performance. The basic paradigm adopted in this and numerous other studies is to present the Selection task in a more or less standard form to naive subjects, expose them to the therapy, then repeat the task and note the changes. In Wason (1968) two such therapies were used, the projection of falsity and the restricted contingency programme, in two experiments. In both cases the task was to select those cards which would enable the subject to find out whether the rule was true or false; it was an 'if then' rule about letters and numbers. The first therapy consisted of asking the subjects which values, when associated with each of those given on the cards ($p$, $\bar{p}$, $q$, $\bar{q}$), would make the sentence false. On repeating the Selection task there was no significant benefit due to this therapy relative to a control group who repeated the task without it: the number of times the $\bar{q}$

card was selected increased by only three, from five to eight. For the restricted contingency programme the subject first evaluated the four truth-table cases (see Table 6) having been given the hint that only one falsified the sentence. All subjects picked out the $p\bar{q}$ instance as the falsifying contingency and pq as the only verifying case, confirming the defective truth-table. However, this made absolutely no difference to their Selection task performance: the response patterns for a group given this experience and a control group were all but identical, and conformed to the pattern noted above.

These results led to further explorations into this apparent conflict between selection and evaluation performance. Two further studies used an evaluation procedure after an initial Selection task (Wason & Johnson-Laird, 1970; Wason & Golding, 1974), both also involving additional manipulations in which "everything was done to encourage the subjects to gain insight" (Wason & Johnson-Laird, 1970). Both presented all the potential information on one side of the cards, using masks to cover what would normally be on the reverse sides, and made even greater efforts, using an interview, to elucidate the role of the critical cards - this was aimed mainly at demonstrating that a p card with $\bar{q}$ on it performed the same (falsifying) function as a $\bar{q}$ card with p on it. Some success resulted: 58% of the subjects in the Wason & Johnson-Laird experiment ultimately made the correct selections, and 35% of Wason & Golding's subjects eventually did so. Presenting all the information on the one side of the cards had no beneficial effects in itself, and neither did Wason & Golding's use of alternative rule forms in which implication was expressed in 'Whenever p, q' sentences, simple assertions, or sentences in which the consequent was mentioned first.

In a similar line of attack, Wason (1969a) tested two more possible sources of difficulty and improvement. In his 1968 experiment the negating values of p and q had been essentially unpredictable in that they could have taken a variety of forms. Perhaps there was a difficulty in hypothesising the form a $\bar{q}$ item could take? Accordingly, item values were made strictly binary: rules were about triangles or circles which could be red or blue, and this restricted universe was made explicit to the subjects. In addition, Wason set out to force a recognition of the inconsistency between evaluations and selections by getting the subjects to contradict themselves in an interview in which the critical cards were discussed. Self-contradiction had been found to improve performance in a thematic inference task (Wason, 1964). Two forms of contradiction were used: one resulting from a discussion of what could be on the other side of the cards, called 'hypothetical contradiction', and 'concrete contradiction' resulting from a revelation and evaluation of what exactly was on the cards. After all this, if the subject still failed to select $\bar{q}$, he was told he was wrong and given a last chance at the task. Frequency of $\bar{q}$ selections increased during the progress of this experiment from an initial zero to 16% after hypothetical contradiction, 31% after concrete contradiction, and 47% after the last chance. These did not just increase the number of correct $p\bar{q}$ selections - there was a large rise in the proportion of $pq\bar{q}$ selection too, from 6% initially to 57% after concrete contradiction.

The truly remarkable thing about these experiments, as Wason and Johnson-Laird have themselves emphasised more than once, is not the improvements in performance so much as the numbers of subjects who never, no matter what is done to them, select p and $\bar{q}$. The impression which emerges from a view of this research is of

experimenters doing everything short of actually placing the subjects' hands on the right cards. If ever there were opportunities for Rosenthal-type experimenter effects, or indeed telepathy, it was surely in these experiments. In a way it is quite heartening to see subjects single-mindedly following their own illogical paths, doggedly resisting attempts to deflect them. Admiration for the independence of the human spirit aside though, we are at this point still left with the question of explaining the observed data, and this is a job for theory. Divergent theoretical approaches have arisen, and for the first of these we turn to the dominant figures in the Selection task story so far.

Insight theories

Johnson-Laird & Wason (1970a) present what may be called the Insight theory of Selection task performance. Wason (1966) laid the foundations for this with his two-factor proposition to account for initial selections: subjects adopt the defective truth-table, where the $p\bar{q}$ and $\bar{p}\bar{q}$ instances are regarded as irrelevant, and so reject the $\bar{p}$ card, and through an overlearned verification habit seek only the remaining instance which proves the rule true, i.e. pq, and select the p and q cards. Subsequently, as we have seen, it was found that reiteration of the task led to a high incidence of initially rare $pq\bar{q}$ selections, and that task performance was at variance with truth-table performance, and so the full Insight model was formulated to embrace all these findings. In its original form the model was presented in two guises with accompanying flow-charts; for the purposes of the present discussion it is presumed that a verbal description of the 'revised' model alone will serve to communicate the important points. The model, then, proposes three levels of insight into the Selection

task: none, partial, and complete. On confronting an 'If p then q' rule subjects are assumed to retrieve the defective truth-table and, lacking insight, focus only on the items named in the rule, selecting only what will prove the rule true. If the subject assumes the rule implies its converse (i.e. is an equivalence), he will select p and q, if not (implication) he selects just p. Going from this to a state of partial insight entails a realisation that one should also see if the rule could be proved wrong, although there is still a need to verify. Thus all those cards which could have verifying and/or falsifying instances on them need to be seen, including items not mentioned in the rule itself, and the subject selects p, q, and $\overline{q}$. On acquiring complete insight the subject realises that he should only be looking for falsifying instances; only p and $\overline{q}$ could falsify, so he selects these. The two stages where insight comes into play are not independent (one of the revisions of the original model) - subjects need to pass through partial insight to get to complete insight. Whether or not the subject attains either state of insight depends on his perception of the cards as reversible, and his realisation of the potential status of the items as truth-table cases. These factors are presumed to be operating in the therapeutic procedures described above, and provide an explanation for the hitherto baffling appearance of $pq\overline{q}$ selections.

As it stands, this model is in immediate trouble, mainly on two counts: imprecision of certain statements and assumptions, and lack of independent empirical confirmation. Two further models have been postulated to account for the former, and these will be dealt with first, as the second criticism can be applied to them as well as the Insight model. Firstly, the weaknesses in the original (revised)

formulation of the model. These centre on the vague references in Johnson-Laird & Wason (1970a) to the role of interpretation. We saw in the previous chapter how there seems to be as much evidence for a defective equivalence interpretation of the conditional as for defective implication (leaving aside the question of response biases), but such interpretational differences are only invoked in the first, 'no insight' stage of the Insight model. In the 'partial' state, Johnson-Laird & Wason state that subjects who did not originally select q will now do so "because it could verify". This is exactly the same reason why they are supposed to select items in the 'no insight' stage, so we are confronted either with an increased verification tendency (along with a new falsification tendency) or a change from implication to equivalence interpretations in the partial insight stage. Neither of these sounds much like insight. Besides, should not equivalence interpreters with complete insight select $p\bar{p}q\bar{q}$? Smalley (1974) attempts to get round this problem by presenting a new model of the Selection task which incorporates the Insight model as the last of three stages. These stages are: interpretation of the rule (defective implication or equivalence), interpretation of the instances (reversible or not), and application of a decision rule to decide what to select in the face of these interpretations. The second stage only partly accounts for another aspect in which the Johnson-Laird & Wason model is vague - the subjects' perception and utilisation of the cards. There is still no precise account of how a subject may proceed from his complete insight that falsification is all that is needed to his actual selection of the appropriate cards, a potent source of difficulty as we have seen; to say that this is part of the package of complete

insight is not enough.

Smalley's model describes 12 possible states for the subject to be in according to his interpretations of the rule (2) and cards (2) and state of insight (3), and he conducted an experiment to test it. At this point his scheme comes unstuck. Firstly he makes little of reversibility of cards in his analysis, except to note, rather mysteriously, that people who did not see the stimuli as reversible would not change their selections during 'therapy'. Why not? Could they not acquire reversibility as part of, or as another form of, insight? More importantly his classification of the types of rule interpretation cannot be considered independent. This is because he used a design similar to that of Wason (1969a) using hypothetical and concrete contradiction; the task was conducted as a group test using written evaluations and comments. In this procedure the evaluation of items as truth-table cases could only take place when all the items were fully revealed - i.e. after the concrete contradiction. With evaluations after two revisions of an initial Selection task and the therapeutic procedures in between, it is hardly surprising that there was a significant relationship between selection and evaluation.

Bree & Coppens (1976) also seize on the role of interpretation in the Johnson-Laird & Wason model. They point out, quite reasonably, that interpretational differences should be reflected not in the "processing considerations" of the Selection task but in truth-tables, in other words that the Insight model is inadequate in accounting for the differences in initial p and pq selections. They present an alternative model which is similar in structure to the Insight model and which shares Smalley's emphasis on interpretation. They propose two possible interpretations (defective implication and equivalence

again) distinguishable by a truth-table task, and three possible Strategies, which are similar in their effects to the three states of insight but are not construed as such. An experiment was run to test this model, using a single Selection task and a truth-table task using fully-revealed examples of the task cards. The report of this experiment is beset by niggling inaccuracies (see Moshman, 1978), but it seems that 19 of the 24 subjects evaluated the rule, according to the truth-table task, as defective implication or equivalence, and 18 of these selected cards in combinations predicted by the model. However, since the only empirical separation of the consequences of this and the Insight model lies in the former's explicit distinction of p and pq selections, it is somewhat unfortunate that only one subject was a p-alone selector. To their credit, Bree & Coppens admit the deflationary effect of this shortage on their model, but set out conditions for a stronger test in future. However, one can criticise this model on other counts: there is nowhere a statement of how or why a subject will adopt a certain strategy, or whether subjects will move from one to another. Bree & Coppens also repeat the mistake of assuming that truth-table tasks constitute a pure measure of interpretation, as if there were no 'processing considerations' involved in them. We saw in the previous chapter that such considerations should indeed be borne in mind when interpreting truth-table data. Thirdly, the model itself seems rather all-encompassing in its rehash of a scheme to account for the commonest selection combinations, and this has led to a more general tilt by Evans (1977b), at the theoretical status of all these insight/strategy models. That is, that they are all framed after the fact - shapes, as it were, drawn around data already collected: states of insight, strategies and interpretations, can

only be deduced from the response patterns they are supposed to underlie, which of course is an exercise in circularity. Now if there were an independent test of any of these states or strategies from which selection data could be predicted before it was seen, then the insight/strategy models' position would be more secure. Such a test has been claimed.

This was the use of subjects' introspections - to find out which strategy a subject is using, why not ask him? Wason & Johnson-Laird mention the correspondence between the parameters of the Insight model and their subjects' verbal reports, as does Smalley (1974). Three particular experiments set about a detailed examination of protocol evidence: those of Wason & Johnson-Laird (1970), Goodwin & Wason (1972), and Wason & Golding (1974). The paradigm used by Wason & Johnson-Laird (1970) and Wason & Golding (1974) is familiar: an initial task, revelation and evaluation of the critical cards, followed by a revision of selections. In most cases, there is a difference between subjects' evaluations of these cards and their treatment of them in the Selection task, and the protocols are enlisted to explore the reasons for this conflict. In Wason & Johnson-Laird (1970) the p card was revealed as having a q item on it and in Wason & Golding (1974) it had a $\bar{q}$ item on it; both experiments used a $\bar{q}$ item hiding a p. There is thus a slight difference between the two studies in the nature of the conflict evoked. In the former, it is in the fact that p can be associated with both verifying and falsifying cases, but in the latter it is in the identity of the p card hiding $\bar{q}$ and the $\bar{q}$ card hiding p. This is of little import where the protocols are concerned though, as the subjects' comments take similar forms: the reasons the subjects gave for their initial choices were almost always consistent with the states in the

Insight model which refer to them, and subjects greeted with inconsistencies between their selections and evaluations tended to say strange things, e.g. they would deny the relevance of the $\bar{q}$ card in the Selection task, or say that if falsified the rule in one task but not the other. Subjects, when pushed, tended to stick by their original selections. In the Goodwin & Wason (1972) experiment the procedure was rather different but the effect much the same; there was no intervening evaluation task, but an experimental group had a set of fully revealed selection cards before them during the task. Both this group and a control group were asked to write down their selection reasons (removing the bias which an interview might entail) and invited to revise their choices, with comments, afterwards. The presence of four fully-exposed cards had no effect, initial selections being much as we have seen them. As with the other two studies, there was support for the Insight model from the protocols - only one was not consistent with it. Seven subjects changed their selections after giving their reasons: five to $p\bar{q}$, one from $p\bar{q}$ to pq, and one from $pq\bar{q}$ to pq. These results look good for the Insight model, but there are some disquieting implications, noted by the experimenters, if one follows it through. What of the inconsistency between selection and evaluation, and the subjects' peculiar commentaries to this? Wason & Johnson-Laird (1970) seem lost for a simple answer; the behaviour of their subjects is not what one would expect of highly intelligent individuals, and they interpret their findings in terms of the Selection task constituting self-instruction of an erroneous solution, which becomes immutable - the selection and evalation processes "pass one another by". This idea of imposing an erroneous structure on the task is continued in Wason & Golding (1974). The 'regression' by some subjects in the Goodwin & Wason study to less

insightful solutions is a more serious matter, resolved by the authors by supposing insight to be liable to fluctuation. This certainly seems the case in an additional test in Wason & Golding (1974), where insight is tested in a transfer selection task: insight gained by nine subjects in one task was only transferred to another by three of them (a similar result was found by Johnson-Laird, Legrenzi, & Legrenzi, 1972).

Wason & Golding admit another possible explanation of the verbal behaviour of Selection task subjects: it could be a rationalisation of already observed behaviour. It is as if the subject, reflecting on the fact that he has selected (say) the verifying cards, explains what he has done by saying that he was trying to verify. Social psychologists of the cognitive-consistency persuasion might be inclined to agree with this account: it accords nicely with the predictions of dissonance theory (Festinger, 1957). Wason & Golding do not like it because "it can hardly account for the acceptance of self-contradictions in the face of what would appear to be undeniable facts" (viz. the identity of the selected $p\bar{q}$ card and the unselected $\bar{q}p$ card). To a dissonance theorist these are just the conditions which would lead to the construction of apparently irrational explanations of the selection-evaluation conflict of the type described. Obviously what is needed is a test of whether the subject is truly reporting states of insight or simply offering explanations of his own conduct.

Dual processes

It has been observed that subjects tend to focus on cards mentioned in the rule (e.g. Johnson-Laird & Wason, 1970a), and Evans & Lynch (1973), in a test of the matching bias hypothesis in the

Selection task (shortly to be enlarged upon), found this to be the case even when negatives were introduced into the rules, changing the logical outcome of a given card selection. For example, the card bearing the named consequent value constitutes a verifying instance of an 'If p then q' rule and a falsifying instance of an 'If p then not q' rule, the logical outcome of the p card being the same for both - verifying or falsifying. Obviously, behaviour based on just matching would be inconsistent with an insight theory, since to select q under the 'If p then not q' rule would mean the subjects were apparently getting the right answer for the wrong reasons, and it is of interest to see what their explanations of such choices would be, or even to assess the effects of such introspection on insight into the task. Perhaps having to give reasons in the negative task would lead to more correct solutions in the normal task?

Wason & Evans (1975) tested these ideas in an experiment, and crucially they gave each subject both kinds of rule. Of the 24 subjects, none chose the correct $p\bar{q}$ combination in the affirmative task and 12 selected pq; this proportion of falsifying and verifying selections was neatly reversed in the negative task, where 15 subjects selected pq, now the falsifying combination, and none selected the now verifying $p\bar{q}$ pair. There was no evidence of an order effect: selecting and commenting on falsifying instances in the negative task did not affect choices in a subsequent affirmative task. Introspections were classified as either mentioning or not mentioning falsification: four did in the affirmative task (it is not clear which selection patterns these were associated with), and 11 did so in the negative task; nine of these subjects selected the falsifying combination, and four of them were the four who mentioned falsification in the other task. Interestingly, only one of the nine did the

negative task second, suggesting some vestige of bias to verify, at least at the introspective level. These findings pose considerable problems for the Insight model. Both the selection frequencies and a significant proportion of the introspections lead to the highly implausible conclusion that subjects may be completely insightful one minute, only to lose it the next - on a rule which ought to be easier to process, being unnegated. It seems far more likely that subjects are simply constructing interpretations of their own behaviour, and if so this diminishes the credibility of protocol data as independent corroboration of internal states such as insight. Wason & Evans (1975) posit a dual process theory to account for these results, "a pessimistic sort of theory" (Erickson & Jones, 1978) which in its strong form implies that behaviour determines thought; if this is so, it applies only to the minority, albeit significant, who performed in this way in the experiment. It is more likely, say the authors, that the unconscious, non-introspectible reasoning process and the conscious verbal process interact with each other to produce the performance observed. The processes can act independently though, as Evans & Wason found in a further study (Evans & Wason, 1976). They used a similar method to that of an early experiment by Wason (1969b), in which subjects were given the problem and the correct solution and asked to explain it. All did so in the correct terms. Wason interpreted this as evidence for insight into the problem resulting from preventing subjects from imposing their own erroneous conctructions on it, but the dual-process results raise the suspicion that the explanations may have been spuriously correct justifications of observed behaviour, in this case someone else's. The obvious test is to present incorrect solutions as correct ones and see if people would agree with and justify them, and this is what Evans & Wason (1976) did. The deception worked,

because most subjects did indeed express agreement with the solution offered and constructed justifications, sometimes of a rather tortuous kind, to support them. However, as the authors concede, this experiment looks as much like a demonstration of social compliance as anything else, since the subjects were simply presented with some supposedly correct solutions to a baffling problem and were under no instructions to doubt them. There is a snag to this idea owing to the use in the experiment of a confidence rating of agreement: most subjects expressed highly confident agreement with the proferred solutions when they could just as easily have maintained agreement and rated their confidence on the low end of the scale. One could of course say that use of this scale was also a matter of compliance, but this would be several steps down the road of construing all behaviour of subjects under instruction as compliance. Nevertheless it is hard to be highly confident about the merits of this experiment as a strong test of the dual process hypothesis - both processes should be involved, especially if it is to be applied in its weaker mutual-feedback form.

Before passing on to the final phases of this review and completing the circle by returning to response biases and content considerations, some points on the embattled Insight approach and the formalistic method of investigation of the Selection task would not be out of order. Wason & Johnson-Laird have voiced repeated concern at the Piagetian view of formal reasoning (to be considered in the Discussion), which, briefly, states that in reflecting on a logical problem the reasoner extracts that problem's logical propositions, recasts them in an abstract and usually conditional form, and subjects then to a combinatorial analysis directed at testing for a falsifying contingency. This is exactly the reverse of what most people do when confronted with the Selection task. However, Piagetian extraction looks

rather like the assumption underlying the work aimed at inducing "insight into a logical relation", namely that there is a logical relation in there which people will recognise when certain obstacles are removed. This relation is unlikely to be material implication, since subjects patently lack it in their repertoire, and besides which the task could be solved correctly by applying the defective truth-table. Perhaps, for most subjects, formal or defective implication is simply not appropriate to activities involving secret properties of cards. An alternative would be to consider insight as referring to a learning process, rather in the Gestalt sense of the word, and to view the Selection task, especially when associated with therapeutic procedures, as a didactic instrument for conditional logic, as hinted at by Wason & Johnson-Laird (1970). Cold water is poured over this argument by Fodor (1976, cited by Johnson-Laird & Wason, 1977), who argues against the notion that growth in the logical capacity of the intellect may arise through sheer experience. Is insight in this paradigm to be conceived as a process of invention then? If it is then what kind of insight or invention is it that is so transient and variable as that which we have observed, which will not transfer between one presentation of the task and a second? We should perhaps lay insight quietly to rest, and consider the Selection task in purely psychological terms; such is the tenor of the next, and final, sections.

Matching bias and stochastic processes

In the previous chapter the recent discovery, or rather systematisation, of an important variable in truth-table task performance was reported: Evans' (1972b) matching bias. It will be recalled that this factor arose from the balanced manipulation of negative components in conditional rules (see Table 4) such that the appearance or non-

appearance of an item in the rule itself and its logical consequences were separated. It was found that subjects, in evaluating the four truth-table contingencies against the four rules, tended to regard the negated, or mismatching, antecedent cases, i.e. $\bar{p}q$ and $\bar{p}\bar{q}$, as irrelevant, whatever their logical consequences. The extension of this idea to the Selction task is fairly straightforward, and has been pre-empted in the discussion of introspective evidence above: using the rotated-negative design and substituting selection cards for truth-table cases, one could separate verification bias, which is central to the Insight model and its progenitor in Wason (1966), from matching bias. If subjects have an overlearned habit of seeking out verifying instances, these should be selected whether or not they are named in a rule, but if subjects are ruling out mismatching cases, they should do so whether or not they verify. Matching and verification are confounded in the double-affirmative rule, so a true item is also an affirming or matching item, while falsifiers also mismatch. In Table 7 the separation of these roles among the four forms of the rules with negatives is given, with the logical values of true/false for the antecedent and consequent balanced against the appearance or not of card items in the rules. The experiment using this design was conducted by Evans & Lynch (1973), testing the matching bias predictions that there should be more selections of cards which affirm, or match, the antecedent and consequent items on rules where the antecedent and consequent are affirmative compared with the corresponding rules where they are negated. Similarly, there should be more selections of falsifying cards when the corresponding rule components are negated. The Insight model would predict a majority of selections, assuming an initial state of no insight, of cards which verify the rule components (TA and TC in Table 5) across all rules. The logically correct

TABLE 7 The relation between the matching and logical values of the items shown on the four cards in Wason's selection task over the four rules with systematically negated components. (After Evans & Lynch, 1973)

| | | Logical case | | | |
|---|---|---|---|---|---|
| | Rule | TA | FA | TC | FC |
| AA | If p then q | p | $\bar{p}$ | q | $\bar{q}$ |
| AN | If p then not q | p | $\bar{p}$ | $\bar{q}$ | q |
| NA | If not p then q | $\bar{p}$ | p | q | $\bar{q}$ |
| NN | If not p then not q | $\bar{p}$ | p | $\bar{q}$ | q |

TA = True Antecedent, FA = False Antecedent, TC = True Consequent, FC = False Consequent. p and q are matching values, $\bar{p}$ and $\bar{q}$ are mismatching values.

solution on all rules is TA-FC, corresponding to $p\bar{q}$ in the ordinary task. All the predictions of matching bias were confirmed in the experiment, and there was no evidence even for a minority tendency to verify, in fact there were more FC than TC selections overall. There were other overall tendencies: TA-FC was the most common combination, and TA and FA frequencies were far less susceptible to matching than were consequent selections.

These results, along with the other points against the Insight model, led to Evans (1977b) undertaking a radical reformulation of explanation of behaviour on the Selection task. He emphasises that, as it stands, Insight provides no explanation of performance, since states of insight are only deducible from the selection behaviour they are supposed to underlie, and in the absence of an independent test this is tautological. Furthermore, the Insight model and its derivatives regard the combinations of card selections as the important index. Evans notes that, throughout the experiments, it is only the incidence of q and $\bar{q}$ (or TC and FC) which varies - p and $\bar{p}$ frequencies hardly change at all. If it could be established that the selection probabilities of individual cards were statistically independent, things would look bad for less parsimonious models emphasising combinations. By reanalysing the selection frequencies of q and $\bar{q}$ against their non-selection in a contingency table, using data from previous experiments, Evans (1977b) found no evidence against the independence of individual card selections. This helps to explain the appearance of $p\bar{q}$ and $pq\bar{q}$ selections in the Wason (1969a) experiment - the combinations are a statistical accident due to the independent increase in $\bar{q}$ selections added to the already common and stable p and pq combinations. Evans therefore proposes a simple probabilistic

model which is based, recalling his 1972c scheme, on a weighted addition of interpretative and operational factors, which in this case correspond to logical tendences, which were not entirely abolished by matching in the Evans & Lynch study, and matching itself. This is the formal statement of the model:

$$Pr(r) = \alpha.I + (1 - \alpha).R$$

Pr(r) is the probability of a reasoning response, I and R the interpretative and operational factors, and $\alpha$ the weighting factor. I, R, and $\alpha$ lie between 0 and 1. The weighting factor is needed to account for instances where the logical tendency is high and the influence of matching close to zero, as in the case of affirmative antecedent rules, so that we are still left with a realistic response probability; simple multiplication would not allow this. The differential effects of matching between antecedent and consequent necessitate different $\alpha$ values for each.

This juggling of parameters to fit the model to already observed effects puts one in mind of criticisms of the Insight model on the ground of circularity. Evans is aware of this, and offers the defence that restating theoretical ideas in mathematical terms directs attention to irregularities in sets of data. However, in the example he cites of the difference between matching effects on the antecedent and consequent in the Evans & Lynch results, he was in fact quite efficient himself at spotting the effect without the aid of mathematics some years before the birth of the model. Furthermore, the model does not answer the question, posed in the 1977b paper, of whether response probabilities are intrinsic to the individual or the group: if, out of 100 people, 60% say 'yes' to a question, is this because each individual is predisposed to say 'yes' 60% of the time, or

because 60 of them always say 'yes' and 40 of them always say 'no'? The Evans model can give no answer. Where it does score over its rivals is in its attempt to describe the interaction of interpretative and operational factors (within individuals), and its success in dealing with the variability of reasoning data, although one could be churlish enough to argue that a verbal formulation of the model could do these things just as well.

Thematic materials

One crucial test of the model which is not inherent in its structure has not been considered yet. The model assumes that I and R values are constant, and that it is only the relative weighting of them which will vary. This weighting could be influenced by the comprehensibility of materials, for instance, or some other determinant of logical or response factors, and one ready instrument for this is the manipulation of the content of the problem. Thematic materials might alter this weighting; there is evidence that they alter the Selection task.

There is a healthy set of data constituting this evidence, appearing in publications over a 5-year period. Apart from negating the consequent of the rule, the use of thematic materials has been the most consistently successful procedure in raising correct response rates on the Selection task. A brief examination of the relevant papers is given here, with a more detailed and critical assessment to come.

In a paper by Wason & Shapiro (1971), two forms of the Selection task were presented to two independent groups of subjects: a normal form with a rule about letters and numbers, and a thematic form with the following kind of rule: 'Every time I go to Manchester

I travel by train'. The four cards, each with a town name on one side and a transport name one the other, showed Manchester (p), Leeds ($\bar{p}$), Train (q), and Car ($\bar{q}$). Two out of 16 subjects picked $p\bar{q}$ in the abstract group, but 10 did so in the thematic group. This significant effect of materials could have been due to a number of factors, and the authors helpfully suggest what they might be. It could have been the meaningfulness of the materials themselves or the connection between them, or the thematic material could have formed a coherent, unified whole so that the subject could more readily transfer attention between the cards, or even away from them altogether. The task could have promoted a greater readiness to entertain alternative, possibly falsifying, hypotheses, or inhibited the imposition of erroneous structures. The 'coherent whole' idea received ample confirmation from a study by Johnson-Laird, Legrenzi & Legrenzi (1972), in which subjects were given the task in the context of a postal sorting job: the rule was 'If a letter is sealed then it has a 50 lire stamp on it' , and the cards were replaced by real envelopes which were correctly stamped, understamped, sealed, or unsealed. In an abstract condition, undergone by the same subjects, envelopes with letters and numbers on either side and a rule about them were used. With 'if then' rules 8.5% of subjects were correct in the abstract task while 87.5% were correct in the thematic task, a startling result and to date the largest recorded effect of thematic materials.

These results were independently supported by Lunzer, Harrison & Davey (1972) using rules about lorries and loads, an additional procedure where only consequent items were presented for selection, and therapies. Two subsequent studies set about the materials v. connection question, untouched by the ones reported so

far, with conflicting results. Gilhooly & Falconer (1974), using four tasks corresponding to the four possible combinations of abstract and thematic terms and relations, found that it was the terms, not the relations, which led to increased $p\bar{q}$ selections. Bracewell & Hidi (1974) used a similar procedure and an additional device: an alternative rule form in which the consequent item preceded the antecedent. This had been found to have some facilitating effect by Wason & Golding (1974), but it did not do so this time. Pooling across the connection and order factors and comparing results between materials, there were 12 (out of a possible 48) correct selections with thematic materials and 9 with abstracts; when the comparison is between connections, the results are more clear-cut: 18 correct with a meaningful connection and 3 with an arbitrary connection. The difference in results between these two studies is mystifying, as the materials used are almost identical. Using different procedures, van Duyne (1974, 1976) obtained more confirmatory evidence for the realism effect. In the earlier experiment logically equivalent linguistic connectives were used to assess the possible role of the actual form of the 'if then' rule. He used 'If p then q', 'Every p is q', 'p or q', and 'not (p and q)' rules and found the materials effect on the first two only. As these were also the rules subjects rated as easiest to understand, it seems that comprehensibility may be a relevant factor. In the later experiment self-generated conditional sentences were used and very high rates of $\bar{q}$ selections achieved, around 90% overall. No comparison with an abstract task was used, but that was not the object of the experiment (more of what was later).

The clear necessity therefore, on the basis of the matching bias results, the evidence for the facilitating effect of thematic

materials, and the theoretical interest due to Evans' (1977b) scheme, is to combine the Evans & Lynch technique with realistic content. We do not know quite what to expect from this; matching may be a prime determinant of responses in the abstract task, but if matching is "a response peculiar to extreme bafflement" (Johnson-Laird & Wason, 1977) resulting from the meaninglessness of the abstract task, one might predict its total disappearance in a realistic context. On the other hand, Evans' stochastic model predicts only a shift in the weighting factor, and this implies a shift towards logical tendencies at the expense of a weaker but still detectable matching effect. Only by varying negative rule components and realism of the problem content can these alternatives be distinguished, and this is the object of the first experiment.

## PART TWO: EXPERIMENTS

### CHAPTER 5



Tables

INTRODUCTION

The object of the first experiment was to test predictions arising from Evans' (1977b) probabilistic model of behaviour on the Selection task against alternative possibilities. Evans' model predicts that increasing the realism of the problem content should shift the weighting of response tendencies away from non-logical biases such as matching bias towards a performance more in accord with the dictates of formal logic. As the shift is in probabilities of responses, it would be expected that detectable patterns due to both interpretative (logical) and operational (matching) tendencies would still emerge. An alternative approach would be to consider the change in reasoning performance brought about by realistic materials to be qualitative rather than quantitative - in line with the evidence from previous studies - and to expect the disappearance of detectable matching responses. These alternatives cannot be distinguished using the normal affirmative rules; abstract and thematic materials in rules with systematically negated rule components are therefore used, as in the experiment by Evans & Lynch (1973) with only abstract materials. This experiment thus offers the opportunity to replicate their findings, as well as investigate the main question.

The abstract materials chosen were of the usual kind, concerning letters and numbers, but it was felt necessary to use thematic materials which have not appeared before in the literature. This was because the ones which have appeared, e.g. in Wason & Shapiro (1971), would be susceptible to aberrant effects due to the presence of negatives. To take an example, negating the antecedent to produce the rule 'Every time I do not go to Manchester I travel by train' results in a nonsensical sentence and an unnecessary distraction. This can be avoided by using rules about food and drink, which are more amenable

to negation while still retaining realism.

## EXPERIMENT 1

### METHOD

#### Subjects

Forty-eight male and female students of Exeter University served as subjects. They undertook the experiment as part of an ancillary psychology class, and were tested as a group.

#### Task and materials

There were two groups of subjects, hereafter called the Abstract and Thematic groups according to the content of the problems they received. Each subject was given a test booklet consisting of an instruction sheet, a sample selection card, and four test sheets, one for each of the four rules (AA, AN, NA, NN). The sample card was included to provide the subjects with a concrete example of the cards referred to in the rules and instructions; it bore items of the kind mentioned in the rules, but these items appeared neither in the rules themselves nor on the test sheets. It was in fact a $\overline{pq}$ card for all the rules. The actual nature of the subjects' task is best communicated through the written instructions given to them.

For the Abstract group they were:

"Thank you for agreeing to participate.

This experiment is concerned with how people reason. Please don't regard it as some kind of intelligence test - it isn't.

You will be given a series of logical rules which may be either true or false. Each rule defines a relationship between capital letters and single-digit numbers in four separate letter-number pairs. Here are some examples of the kinds of rule referred to:

'If the letter is an N then the number is not a 3'

'If the letter is a T then the number is an 8'

'If the letter is not a G then the number is a 7'

'If the letter is not a C then the number is not a 6'

Each answer sheet has a rule of this kind at the top. Below each rule is a picture of four cards, each of which represents a particular letter-number pair. Each card has a letter written on one side and a number on the other side, but naturally, only one side of each card is visible in the picture.

The logical problem is to decide, for each rule, which of the cards would need to be turned over in order to find out whether the rule has been obeyed or not.

You are free to choose any or all of the cards as you think necessary. An example of the kind of card which the pictures refer to is attached. Take your time over solving these problems: quite often they are not as easy as they appear at first. You may refer back to these instructions whenever you like".

For the Thematic group the relevant parts of these instructions were changed so that the relationship defined by the rule was "between what I eat and drink together at separate meals", the picture of cards representing "what I ate and drank at a particular meal. Each card has what I ate written on one side and what I drank written on the other side..." Four examples of food and drink rules were given:

'If I eat fish then I do not drink beer'

'If I do not eat chicken then I do not drink wine'

'If I eat pork then I drink whisky'

'If I do not eat potatoes then I drink tea'

Each of the test sheets in the booklet bore a rule at the top, four cards' drawn below, and the additional instructions: "Please indicate which of the cards drawn below would need to be turned over to find out whether the rule has been obeyed or not. Please tick (✓) any of the cards you think would need to be turned over, and cross ( X ) any which you think would not need to be turned over. Please do not leave any unmarked". A sample test sheet appears in Appendix A.

Thus the subjects had to make positive decisions to select or reject the cards. None of the rules given in the instructions to either group appeared in the task itself. For both groups the allocation of items to rules and cards was partially randomised - no item appeared in more than one rule in each booklet. The food and drink items were taken from a set of eight of each, prepared in advance, and the letters I and O and the numbers 1 and 0 were not used because of possible confusion.

Procedure

The group was given an additional briefing session in which the form, but not the content, of the coming test was explained to them; they were asked to delay their questions until after the test. A full explanation was then given. For the test itself, the group was divided into two halves. The subjects were sat well apart from each other and asked not to confer. The test booklets were then distributed to them.

RESULTS

The frequencies of selection of each card under each rule are recorded in Table 8a. On a casual inspection of the data, there

TABLE 8a Selection frequencies of the individual 'cards' on each rule in Experiment 1. A = Abstract group, T = Thematic group, otherwise notation as in Table 7. N = 24 in each cell.

| | | Logical case | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | TA | | FA | | TC | | FC | |
| | Rule | A | T | A | T | A | T | A | T |
| AA | If p then q | 23 | 23 | 3 | 4 | 15 | 10 | 10 | 7 |
| AN | If p then not q | 24 | 22 | 2 | 2 | 5 | 3 | 18 | 18 |
| NA | If not p then q | 19 | 21 | 7 | 5 | 17 | 19 | 12 | 9 |
| NN | If not p then not q | 20 | 18 | 5 | 8 | 13 | 9 | 16 | 15 |
| | % | 90 | 88 | 18 | 20 | 52 | 43 | 56 | 47 |

Table 8b Frequencies of selection combinations, ordered in terms of logical case, on the four rules in Experiment 1.

| | Rule | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AA | | AN | | NA | | NN | | Overall | |
| Combination | A | T | A | T | A | T | A | T | A | T |
| TA | 5 | 7 | 4 | 5 | 1 | 2 | 1 | 2 | 11 | 16 |
| TA-TC | 8 | 8 | 1 | 0 | 5 | 7 | 6 | 7 | 20 | 22 |
| TA-TC-FC | 4 | 1 | 2 | 1 | 4 | 5 | 6 | 0 | 16 | 7 |
| TA-FC | 3 | 4 | 15 | 14 | 6 | 3 | 5 | 7 | 29 | 28 |
| Others | 2 | 4 | 2 | 4 | 8 | 7 | 6 | 8 | 18 | 25 |

appears to be little difference between the results for the two groups, and this observation is supported in the analysis. Both groups' data were tested according to the predictions of matching bias, as set out in the Evans & Lynch paper (see Chapter 4). The notation used in that paper and in the discussion of it in the previous chapter is retained here, with TA, FA, TC, and FC being used for the logical cases and p, $\bar{p}$, q, and $\bar{q}$ referring to the matching values (see Table 7). It will be recalled that four independent predictions can be made on the basis of matching bias:

(i) More TA selections on rules with affirmative antecedents;

(ii) More FA selections on rules with negative antecedents;

(iii) More TC selections on rules with affirmative consequents;

(iv) More FC selections on rules with negative consequents.

These four predictions were assessed, using sign tests (one-tailed), for both Abstract and Thematic groups, and the results are shown in Table 9. All eight tests were in the predicted directions, although predictions (i) and (ii) in the Thematic group fell just short of significance.

Evans (1977b) in his stochastic model paper showed that card selections should be considered statistically independent, and the above analysis complies with this. Some examination of combinations of selections is merited nonetheless, in view of the possibility that they might reveal different solution patterns within the frequencies themselves. This examination is doubly justified, apart from this possibility, firstly because of the prediction of the Insight theories that TA-TC should be the most common overall Abstract selection and TA-FC the most common Thematic selection, and secondly because Evans has proposed that the use of thematic materials may affect the independence of card selections. Accordingly, Table 8b shows the

TABLE 9 Results of analysis of the four matching bias predictions on data from Experiment 1. Analysis was by one-tailed sign tests

| | Prediction | Group: Abstract | Group: Thematic |
|---|---|---|---|
| (i) | More TA on rules with Aff. Antecedents | $p = .008$ | $p = .109$ |
| (ii) | More FA on rules with Neg. Antecedents | $p = .020$ | $p = .090$ |
| (iii) | More TC on rules with Aff. Consequents | $p < .001$ | $p < .001$ |
| (iv) | More FC on rules with Neg. Consequents | $p = .011$ | $p < .001$ |

relative frequency of the four common combinations (cf. Chapter 4) and the residual 'others' category for both groups under all rules. On this analysis one can see that there is no support for the Insight predictions in either group: neither TA-TC nor TA-FC claims a significant majority, and the combinations share with the individual selections the striking similarity between the two groups. There is no difference between TC and FC selections in either group, and therefore no evidence for minority verification or falsification tendencies. The prospects of both qualitative and quantitative shifts in performance between Abstract and Thematic groups are unfulfilled in this analysis.

## DISCUSSION

These results are both interesting and unexpected. They are novel inasmuch as there is a break with precedent: the use of thematic materials has made scarcely any difference to the subjects' performance on the Selection task relative to that found with abstract materials. Such small points of difference as there are between the groups do not aid an explanation of what has happened. There is a suggestion that the effects of matching bias on antecedent selections, already attenuated by logic in the abstract task, are reduced further in the thematic presentation. There is also an appreciably lower incidence of TA-TC-FC selections in the Thematic group than in the Abstract group; these are replaced by more 'others', which on closer examination consist largely of combinations containing the double-matching (pq) pair on the NA and NN rules, where it is not represented in the common combinations. On the one hand therefore there seems to be less matching in the Thematic group, on the other more, which is not very helpful. These marginal effects merit replication, in the light (i) of the some-

what unexpected findings of the total generalisation of matching bias to thematic rules, and the corresponding (logically) low level of performance in the Thematic group; and (ii) of the major procedural novelties which may have contributed to the discrepancy between these results and those of previous research. Let us consider these differences.

First of all there are differences relating to the conduct of the experiment: the instructions, the test format, and the linguistic content of the thematic rules. Regarding the instructions, it could be that their wording was in some way misleading or insufficient. This would not on any count be expected to affect the responses of the Abstract group, since their behaviour already seems mostly determined by non-logical factors (Evans, 1972c; Evans & Lynch, 1973). However, instructional shortcomings might serve to bring down performance on the thematic task to a similar non-logical level, either by obscuring the requirements of the task or by increasing the bafflement considered by Wason (1977a) to be a likely cause of matching effects. Such shortcomings are not obvious in the present instructions, though, in comparison with those reported in other papers, and no ready differences between instructions then and now present themselves. Perhaps then the mode of presentation of the instructions, on a typed, duplicated sheet which the subjects were asked to read, might have had some effect. Subjects may have given them a less than thorough reading, glossing over points which might receive some emphasis in a verbal presentation (although they meet them again, four times, in the test sheets), and this again would be more likely to affect the Thematic group, for

similar reasons as above. This criticism, if valid, would also have to apply to other experiments using a similar procedure (e.g. Smalley, 1974; Evans & Wason, 1976).

The actual task content could also be construed as an area of difference between this and previous experiments. This does not refer to the abstract materials; these were deliberately designed for close correspondence with those used in other studies, but the thematic materials were equally deliberately different, for the reasons given in the introduction: they refer to food and drink and not towns and transport, envelopes and stamps, or lorries and loads. For the actual thematic materials to have the apparently debilitating effects observed, it would have to be argued that they are in some way less thematic than the ones used before, but if they are it is by no means clear how they are. A weaker version of this proposal, voiced privately to the writer, is that food and drink materials are not quite as meaningful as travel or postal contents, presumably because they do not evoke such an aura of plausible context. The reply to this is twofold: firstly, it is beside the point whether jorneys or meals are the more realistic, since the food and drink rules are still indisputably more meaningful than rules about letters and numbers, so the lowering of correct solution levels (or, more properly, FC selection rates) to that normally only associated with abstract tasks cannot be attributed wholly to such a difference. Secondly, the context in which the thematic materials were framed, i.e. rules concerning separate meals and cards depicting actual consumption, was spelt out explicitly in the instructions, and sounds no more or less coherent than imaginary journeys. A final test of the comparison between different materials calls for replication studies using both, and this call is answered later in the present series.

As with mode of instruction, presentation of materials could also have had a bad effect: the test booklet contained only drawings of cards and this may have hindered appreciation of the reversibility of selection cards, crucial to the solution of the problem. The importance of acknowledging the potential hidden items has been emphasised by other authors (e.g. Bracewell, 1974; Smalley, 1974). It should be remembered that a sample card was given to the subjects as a concrete example of the cards referred to in the test. However, as this was a $\overline{pq}$ card it may have been regarded as irrelevant; in any case, a replication study using actual instead of drawn cards would resolve this particular issue.

A less obvious or well definable set of possible contributors to the unusual results could be cognitive factors inherent in this experiment. The test reported here consisted of four selection tasks, three involving negated rule components. Most subjects in previous Selection task experiments using thematic materials have had to do only one task, and none have faced negated components. This extra task load coupled with the increased interpretational difficulty due to negatives gives rise to the possibility of some kind of cognitive overload. Not only may the greater difficulty of the whole task defeat the subjects, but their perception of the task may be different from that in the usual abstract-thematic experiment, and if this perception is in the direction of heightened difficulty in an already difficult problem, a 'regression' by Thematic subjects to the non-logical response patterns seen in the abstract task might result. This being so, a simplified form of the experiment, omitting the three negative-containing rules, should give rise to the usual facilitation by thematic materials.

Lastly, we come to a major set of factors in which this experiment differs from its predecessors: social factors. These encompass (i) factors arising from the presence of the experimenter, and (ii) factors arising from the presence of other subjects. The first of these is allied to the discussion of presentation of instructions above, and mainly concerns the difference between procedures involving face-to-face, experimenter-subject running and the group testing used here. Although the Selection task has been given as a group test before, only one of the studies using thematic materials has done so (Bracewell & Hidi, 1974) and it employed small groups and a considerably more 'familiar' procedure than did the present experiment. In the main, Selection tasks are normally given to subjects by an experimenter who reads instructions and may record responses. This allows the opportunity for emphasis of salient features of the instructions and materials, repetition, making sure the subject fully understands what is required of him, possibly even to the extent of influencing his decisions. In Experiment 1, these points are all in the subjects' own control: given the instructions and materials, it is up to each individual what to make of them. A face-to-face, verbally instructed task presentation would show whether these points could have influenced the outcome of the experiment. The second social factor relates to the possible effects on subjects' performance of the presence of other subjects - in other words, social facilitation. The specific effect alluded to here is coaction, defined by Zajonc (1965) as "individuals all simultaneously engaged in the same activity in full view of one another" - exactly the procedure used in this experiment. Zajonc reports evidence which indicates that this could lead to performance decrements in tasks such as the Selection task: Allport's (1920) classic

study records that, while coaction appeared to facilitate performance on some simple motor-oriented tasks, the reverse seemed to occur on more complicated problem-solving tasks, with performance in the presence of others being reduced relative to a solo effort. This effect was attributed to coaction increasing the probability of dominant responses, which in complex problem-solving tasks are usually incorrect ones (otherwise there is no problem), over others. There is a relation between these findings and the present experiment: here we have a group of subjects engaged together in a task for which there is ample reason to suspect that the dominant response is an incorrect one. The clear priority then is to examine the specific question of whether a dominant response - matching - is asserting itself over logical responses in the Selection task because of a coaction effect. It was therefore decided to repeat the previous experiment with one modification: subjects would perform the task in rooms on their own, isolated from the presence of an experimenter or other subjects.

## EXPERIMENT 2

### METHOD

#### Subjects

Forty-eight male and female students of Plymouth Polytechnic served as subjects. They were recruited as paid volunteers and tested individually; none reported any previous experience with tasks of this type.

#### Procedure

The task and materials were as in Experiment 1. The difference in this experiment was purely procedural - subjects were tested as individuals, not as a group. Each subject was conducted into a small

room or cubicle, handed a test booklet, told that all the necessary instructions were on the sheet provided, and assured that a full explanation of the task would be given on its completion. Subjects were then left alone to complete the task, and several could be run simultaneously in this way. They were allocated alternately to the Abstract and Thematic groups as they arrived.

## RESULTS AND DISCUSSION

The same analyses were performed as in the previous experiment. Table 10a shows individual card selection frequencies for each group, and again there is an immediate impression of similarity both between the two groups' data and these results and those in Table 8a. That impression is confirmed by inspection of the table of selection combinations (Table 10b), which closely resembles the previous table for Experiment 1. The high number of 'others' on the NA and NN rules is more striking here, but in both experiments the great majority of these is made up of combinations containing the double-matching pair (FA-TC in the NA rule and FA-FC in the NN rule; see Table 7) which are not represented in the 'common' combinations, as these do not include FA.

The results of the four matching bias tests for each group are shown in Table 11. Again, all are in the predicted direction, with one (prediction (i), Abstract group).just short of significance. The effects of matching bias and the lack of facilitation due to thematic materials are strongly re-affirmed: there is no difference between the groups, both performing at the level normally associated with the abstract task.

It therefore seems safe to conclude that there was no

TABLE 10a Selection frequencies of the individual 'cards' on each rule in Experiment 2. Details of notation and N are as in Table 8.

| | | Logical cases | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | TA | | FA | | TC | | FC | |
| | Rule | A | T | A | T | A | T | A | T |
| AA | If p then q | 20 | 21 | 4 | 2 | 16 | 15 | 6 | 7 |
| AN | If p then not q | 23 | 21 | 1 | 4 | 8 | 2 | 18 | 16 |
| NA | If not p then q | 19 | 14 | 8 | 12 | 19 | 22 | 7 | 5 |
| NN | If not p then not q | 17 | 14 | 8 | 14 | 9 | 9 | 14 | 17 |
| | % | 82 | 73 | 22 | 33 | 54 | 50 | 47 | 47 |

TABLE 10b Frequencies of selection combinations in Experiment 2

| | Rule | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AA | | AN | | NA | | NN | | Overall | |
| Combination | A | T | A | T | A | T | A | T | A | T |
| TA | 4 | 5 | 5 | 5 | 3 | 1 | 2 | 2 | 14 | 13 |
| TA-TC | 12 | 11 | 1 | 0 | 9 | 6 | 5 | 4 | 27 | 21 |
| TA-TC-FC | 1 | 2 | 6 | 0 | 2 | 2 | 4 | 2 | 11 | 6 |
| TA-FC | 1 | 2 | 10 | 13 | 1 | 0 | 3 | 2 | 15 | 17 |
| Others | 6 | 4 | 2 | 6 | 9 | 15 | 10 | 14 | 27 | 39 |

TABLE 11 Results of the analysis of the four matching bias predictions on data from Experiment 2

| | Prediction | Group: Abstract | Group: Thematic |
|---|---|---|---|
| (i) | More TA on rules with Aff. Antecedents | $p = .062$ | $p = .011$ |
| (ii) | More FA on rules with Neg. Antecedents | $p = .046$ | $p = .004$ |
| (iii) | More TC on rules with Aff. Consequents | $p = .004$ | $p < .001$ |
| (iv) | More FC on rules with Neg. Consequents | $p = .001$ | $p = .001$ |

coaction effect operating in these experiments - the unusual results cannot be attributed to performance of the task in the presence either of the experimenter or of other subjects, since both were absent in the second experiment. The possibility of the influence of cognitive factors therefore demands attention, especially those relating to task difficulty, either perceived due to the presence of three additional tasks, or actual due to the presence of negation in those three. Both of these possible variables would be eliminated by simplifying the procedure to use only the normal, AA rule in a single presentation. This is done in the next experiment.

## EXPERIMENT 3

### METHOD

#### Subjects

Thirty-two male and female students of Plymouth Polytechnic, recruited on a paid volunteer basis and unfamiliar with the task, served as subjects.

#### Task and Materials

Similar rules to those used in Experiments 1 and 2 were composed, but only the AA form was used this time. As the subjects were only required to carry out the one task, the instructions were amended accordingly. The test test booklet was therefore a truncated but similar version of that used previously.

#### Procedure

Subjects were again allocated alternately to either group as they arrived, and handed the test booklet. As the presence of others has been shown not to influence performance on this task, the subjects were run in small unsupervised groups of three or four, under

instruction not to confer. A full debrief was given after testing, as before.

RESULTS AND DISCUSSION

Table 12 gives the frequencies of individual card selections and combinations. In both cases the striking similarity between both groups' performance is confirmed, as is the lack of facilitation by thematic materials. There is neither an increased tendency to select FC nor a greater number of TA-FC combinations in the Thematic group. There actually seems to be more matching behaviour in this group, inasmuch as this is reflected in the number of TA-TC (p and q) selections. Individual analyses of card selection frequencies for both groups were performed, using the Fisher Exact test. This involved setting the selection and non-selection proportions for each card in a 2X 2 contingency table. The test was one-tailed for the FC ($\overline{q}$) card, because results from research prior to this would lead to a prediction of more FC's in the Thematic group, and the other three tests were two-tailed in the absence of any such a priori prediction. All these comparisons were non-significant - there was no difference between the groups in selection of any card.

It therefore seems that the unusual results in the previous experiments were not due to the presence of extra tasks or negation. This is important, as the procedure in Experiment 3 is closer to that used in most other studies of the Selection task with thematic materials. Of course, one might object that running the subjects in small groups without supervision was an open invitation to cheat, even though they were asked not to. This is both an unwarranted slur on the good name of honest subjects and an implausible explanation,

TABLE 12 (a) Number of times each individual card was selected in each group in Experiment 3. N = 16 in each cell.

| Card | Group: Abstract | Group: Thematic |
|---|---|---|
| TA | 14 | 13 |
| FA | 6 | 1 |
| TC | 9 | 14 |
| FC | 6 | 5 |

(b) Frequency of various selection combinations. N = 16 in each group

| Combination | Group: Abstract | Group: Thematic |
|---|---|---|
| TA | 3 | 0 |
| TA-TC | 5 | 9 |
| TA-TC-FC | 1 | 3 |
| TA-FC | 1 | 0 |
| Others | 6 | 4 |

since both groups' results were consistent with patterns previously found on the abstract task and not the thematic version. Conferring should surely lead to more logical solutions, not fewer. Is is more likely that the disparity between these results and previous must lie in two further variables as yet unexplored: mode of presentation and actual content. All three experiments so far reported have given the Selection task as a pencil-and-paper test, with representations of cards and written instructions, whereas almost all previous studies have used the face-to-face, verbal mode of presentation, with all the possible influences due to emphasis, reiteration, and nonverbal cueing which such procedures inevitably risk. This is therefore the procedure adopted in the next experiment, and the format used is based on that described in the Wason & Shapiro (1971) paper.

## EXPERIMENT 4

### METHOD

#### Subjects

Thirty-two students and technicians from Plymouth Polytechnic were recruited on a paid volunteer basis and tested individually; none had any experience with this type of task.

#### Task and Materials

These were different from the ones used before in certain important details. The task was a standard Selection task and the materials were as before, but this time the selection items were actual cards taken from a deck of 16 which the subject had examined, and the task was to indicate by pointing the cards it was necessary to see to test the rule. All cards bore either a food word on one side and a drink word on the other (Thematic group), or a letter on one

side and a number on the other (Abstract group).

Procedure

An entirely different procedure from that used in the foregoing experiments was used. The task was verbally based, with the experimenter sitting opposite the subject instructing him and eliciting and recording his responses. These were the instructions read out at the start of the testing session:

"Thank you for agreeing to participate.
This experiment is concerned with how people reason. Please don't regard it as an intelligence test, because it isn't one.
Here are some cards; please have a look through them. On each card there is a food word/letter on one side and a drink word/number on the other side".

The subject was then handed the familiarisation deck. After he had looked through it it was taken back and the four test cards extracted, away from the subjects' view. In the Thematic group each subject received one of four different sets of test cards for one of four different rules which were used, to control for possible preconceptions about certain food-drink combinations. The same rule and cards were used for the Abstract group. Slightly different instructions for the two groups then followed. For the Thematic group these were:

"I will now present you with four cards; the test you are about to do concerns only these four cards. These cards show what I ate and drank at each of four separate meals on four separate days, with what I ate on one side of the card and what I drank on the other side. I am going to make a claim about what I ate and drank, and your task is to say which of the cards would need to be turned over to

decide whether that claim is true or false. You may choose any or all of the cards".

The sentences which were changed for the Abstract group reads as follows:

"...These cards show a letter on one side and a number on the other side. I am going to make a claim about the connection between letters and numbers and your task..."

The instructions were repeated if the subject asked for them to be, and then the four cards were laid out in random order on the table. The rule was read out, and the subject asked to take his time and indicate his choice. If he indicated only one card, the experimenter, affecting a casual tone, said "Just that one?", but otherwise there was no further instruction, and any request for more was refused. After testing a full debrief was given.

## RESULTS AND DISCUSSION

Table 13 records the selection frequencies for Experiment 4 in terms of individual cards and combinations, as in Experiment 3. Once again the identity of the two groups is manifest, as is the relation between these results and previous findings on the abstract task. Fisher Exact tests of the individual card selections again showed no difference between the groups. Another variable is therefore accounted for: there was no effect due to a face-to-face mode of presentation, which is perhaps just as well for the validity of the Selection task. There is thus only one other area of difference between these experiments which do not show a materials effect and the ones reviewed in Chapter 4 which do: the nature of the thematic materials. The most popular thematic materials in others' experiments

TABLE 13 (a) Number of times each individual card was selected in each group in Experiment 4. N = 16 in each cell.

| Card | Group: Abstract | Group: Thematic |
|---|---|---|
| TA | 11 | 10 |
| FA | 0 | 3 |
| TC | 10 | 7 |
| FC | 2 | 3 |

(b) Frequency of various selection combinations. N = 16 in each group.

| Combination | Group: Abstract | Group: Thematic |
|---|---|---|
| TA | 3 | 6 |
| TA-TC | 6 | 4 |
| TA-TC-FC | 0 | 0 |
| TA-FC | 2 | 0 |
| Others | 5 | 6 |

have been those first used by Wason & Shapiro (1971): the town and transport rules. By substituting these for the food and drink materials used up to now, and making certain detail changes to the procedure of Experiment 4, which was adapted from Wason & Shapiro (1971), a straight replication of the latter experiment should be possible, and should resolve the question of any linguistic effects. This is the object of the next experiment.

## EXPERIMENT 5

### METHOD

#### Subjects

Thirty-two male and female students of Plymouth Polytechnic, recruited as paid volunteers, tested individually, and having no previous experience with this type of task, served as subjects.

#### Task and Materials

These were essentially the same as in Experiment 4, except that the thematic materials concerned journeys rather than meals.

#### Procedure

There were minor procedural changes required for a close replication of the Wason & Shapiro experiment: in addition to a task instruction about destinations and means of transport, four cards with names of different days were also used. The selection cards were each placed on one of these, showing that the items on the cards pertained to journeys on different particular days. As in Experiment 4, four versions of the thematic task were presented.

### RESULTS

The results are presented and were analysed in the same way

as in the previous two experiments. In Table 14 are the selection frequencies of individual cards and combinations, and the individual selections were again analysed by the Fisher Exact Probability test. There were no differences between the groups, and inspection of Table 14 will confirm the close correspondence with the results found before: both groups performed at the level previously only associated with the abstract task.

## SUMMARY AND DISCUSSION OF EXPERIMENTS 1 - 5

The results from all these experiments appear rather negative, but it would be a mistake to view them so. Firstly, the Abstract data of Experiments 1 and 2 constitute an important double replication of the findings of Evans & Lynch (1973). This replication poses more problems for the Insight theories and their proposition of a general verification tendency, and supports the conclusion that "it is matching rather than verifying which appears to be the main determinant of subjects' selections" (Evans & Lynch, 1973). Not the only determinant though: the strong tendency for antecedent selections to follow the requirements of logic is confirmed as well, by the overall preference for TA over FA whatever the matching value of the relevant item. Pooling the data from the 96 subjects run in Experiment 1 and 2, TA was selected 83% of the time and FA 23%. However, Evans & Lynch's findings of a significant minority tendency to falsify is unsupported here: the proportions of TC and FC selections are 50% and 49% respectively, and there should be more FC selections for a falsification tendency to emerge. The selections of TC and FC are not random: inspection of Tables 8 and 10 show that they closely follow matching, i.e. consequent items tend to be selected when

TABLE 14 (a) Number of times each individual card was selected in each group in Experiment 5. N = 16 in each cell.

| Card | Group: Abstract | Group: Thematic |
|---|---|---|
| TA | 14 | 14 |
| FA | 2 | 0 |
| TC | 13 | 10 |
| FC | 1 | 2 |

(b) Frequency of various selection combinations. N = 16 in each group.

| Combination | Group: Abstract | Group: Thematic |
|---|---|---|
| TA | 2 | 4 |
| TA-TC | 11 | 7 |
| TA-TC-FC | 0 | 1 |
| TA-FC | 0 | 1 |
| Others | 3 | 3 |

they are mentioned in the rule and ignored when they are not. Logic only seems to exert any influence on the selection of antecedent items. This concurs with the findings on the directionality of the conditional from many different quarters (e.g. Evans, 1977a; Evans & Newstead, 1977; Braine, 1978; see Chapters 2 - 4) and is incorporated in Evans' (1977b) model in the differential weighting factor.

The real surprise in these experiments is the difference, or rather the lack of difference, between the materials groups, and it is plain that the initial predictions of a qualitative or quantitative shift in the balance between logic and matching are both unfulfilled.

However, there is some support for Evans' suggestion that realism may affect the statistical independence of consequent selections he found in his 1977b paper. A similar analysis was carried out on the consequent selection frequencies for both groups in all five of the experiments here; viz. the relative selection and omission of the consequent items was ordered in 2 X 2 contingency tables and analysed by the Fisher Exact test. The results of this analysis are summarised in Table 15, and we can see that the statistical independence of consequent selections is maintained throughout in the Abstract groups - selection of one card does not depend on selection or omission of the other. One of the comparisons approaches significance (NA, Experiment 1), but this tendency is not replicated, neither is it suggested in the Evans & Lynch (1973) data (see Evans, 1977b), so no importance can be attached to it. In the thematic data the consistency is disturbed by a significant deviation from independence on the NN rule in Experiment 1 and a similar though non-significant trend in the same place in Experiment 2. On examining the contingency tables it was found that in both cases, selection of one card entailed

TABLE 15 Results of contingency table analysis to test the independence of consequent selection in Expteriments 1 - 5. P values refer to two-tailed probabilities derived from the Fisher Exact Probability tes.

| | Experiment 1 | | | Experiment 2 | | | Experiments 3 - 5 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Abstract | Thematic | | Abstract | Thematic | | Abstract | Thematic | |
| | p | | | p | | | p | | |
| AA | 1.0 | .712 | AA | .604 | .414 | E.3 | 1.0 | 1.0 | |
| AN | 1.0 | .806 | AN | .638 | 1.0 | E.4 | .250 | 1.0 | |
| NA | .068 | .508 | NA | 1.0 | 1.0 | E.5 | 1.0 | 1.0 | |
| NN | .311 | .006 | NN | .612 | .084 | | | | |

A p value below .05 denotes a significant deviation from independence

omission of the other, the FC (q) card being selected about twice as often as the TC ($\bar{q}$) card: 26 subjects selected FC alone, while 12 selected TC. Why there should be a significant bias towards single consequent selections on this rule and no other, and with thematic rather than abstract materials, is not clear. Wason & Johnson-Laird (1972) suggest that conditionals with negated antecedents are interpreted as disjunctions, and that "in everyday life, context would completely clarify whether or not the conditional is to be treated as a disjunction" (p. 62). It is possible that the subjects are interpreting the NN thematic rule as a disjunctive here, which might lead to the selection of just one antecedent or one consequent value. This is far from being a satisfactory explanation however, since we are still left with the question of why NA rules are not treated in this way. This result is a peculiar one, and eludes explanation.

As may be divined from some of these comments, and from the discussions of the individual experiments preceding, the findings of the present research were genuinely surprising, and each exploration of various possible factors contributing to the results of Experiment 1 may be looked on as 'therapies' with the same object as those in some of the experiments reviewed in Chapter 4. That is, that having started with an apparent breakdown in people's performance (Experiment 1), the rest of the series was directed not only at exploring possible variables intervening between these experiments and others, but at progressing steadily towards a set of conditions under which the expected result would emerge. This progress eventually arrived at the point of exact replication of a previous 'successful' experiment, and as the expected result, of the Thematic group performing more

logically than the Abstract group, remained elusive, we are left with two last possibilities. One is that other factors, as yet unmentioned and unexplored, are still preventing the true situation from emerging. Some suggestions have been made as to what these might be; none of them stand up. For instance, mention has been made of population differences between this and previous studies having caused the disparity, but the five experiments here ranged over university and polytechnic populations, so this is unlikely. Experimenter effects have also staked their claim, but four different experimental procedures were involved in this series: supervised groups, unsupervised groups, unsupervised solo, and face-to-face solo. Only in the first and last was the experimenter even in the same room. Even granting that there could still be such undetected influences in the present study, they must be of extreme subtlety, so much so that one would have to question the power of the thematic materials effect. This anticipates the second possible reconciliation of the present results with previous: the thematic materials effect might not be what it seems.

The thematic materials effect has attained an almost lawful status in some of the literature, and was taken as an a priori assumption at the outset of the research reported here. To answer the question of whether it deserves its standing, we must go back to the previous publications and examine them more closely. The description of these papers in the Review section (Chapter 4) was deliberately superficial for this reason; detailed examination of these reports reveals that a reliance on a simple thematic materials effect could only have come from such a superficial reading. The previous papers break down into two main factions: those in which the thematic materials effect is not of the order normally assumed, and those in which additional

variables besides thematic materials per se are involved.

In the first category fall the closely related studies by Bracewell & Hidi (1974) and Gilhooly & Falconer (1974), both of which aimed to investigate the relationship between the meaningfulness of the terms in the rules and the connections between them. They used town and transport and letter-number rules in the contexts of journeys and two-sided cards. As an illustration of the mixture of abstract and thematic terms and connections, consider an abstract terms/concrete relations rule from Gilhooly & Falconer: 'Every time I go to D, I travel by 3'. Bracewell & Hidi, in addition to this, also investigated the effect of putting the consequent before the antecedent, so their subjects, who were run individually in independent groups, encountered one of eight types of rule compared with Gilhooly & Falconer's subjects, also in independent groups, who faced one of the four more usual rules. It will be recalled from Chapter 4 that Bracewell & Hidi found no significant difference in correct responding due to the type of materials: pooled over the relationship and order factors, 25% of subjects confronted with thematic materials gave the correct combination, while 19% of subjects with abstract materials did so. Rather, it was the difference between the connections between the terms which turned out to be significant: 38% of subjects reasoning about rules where the connection was "natural" chose TA-FC ($p\bar{q}$) while only 6% did so when the relationship was "arbitrary". The authors themselves point out that one group was responsible for this difference: 9 out of the 12 subjects in the concrete materials-natural relationship-first order group were correct. This rule of course corresponds to a normal thematic rule, and the abstract materials-arbitrary relationship-first order type similarly corresponds to a normal abstract rule. The comparison of

correctness between these two groups is 75% : 8% respectively. Such a comparison is a double-edged weapon however: when wholly thematic rules have their order reversed, e.g. 'I travel by car every time I go to Ottawa', the effect evaporates, and only two of the 12 subjects get the task right. There is no ready explanation for this - reversing the order of terms has been posited as a facilitator (cf. Wason & Golding, 1974; Johnson-Laird et al., 1972), yet here it is doing the opposite. Bracewell & Hidi offer a tentative explanation of this discrepant effect in terms of the second-order sentences expressing aberrant goal-means relationships, assuming that sentences should be easier to deal with when they mention goal and means in their correct temporal order. The problem with this idea is that the goals and the means in the thematic sentences could equally well be seen as being in the correct sequence in either order: certainly one might first decide the destination before setting out on a journey (first order), but it is also true to say that one must travel before arriving (second order). All in all, this unexpected order factor between the thematic sentences is rather embarrassing for the proponents of a simple thematic materials effect. In addition, Bracewell & Hidi specified to their subjects that the rules were not to be interpreted as implying their converses, and the heeding of this injunction is reflected in the extradordinarily low overall rate of TA-TC (pq) selections - 3%. Surely these efforts to promote an understanding of conditional logic should have had more of a facilitating effect, especially when allied with realism, than the relatively low levels of correct responding shown to have been going on here imply.

The picture is further complicated by the findings of Gilhooly & Falconer's contemporary study of the terms v. relations question.

Using a similar procedure, they procured rather different results. This time, 21% of subjects using rules with thematic terms were correct compared with 9% using abstract. The comparison between relations was 17% correct with concrete and 13% correct with abstract. A significant effect was found due to the terms alone. The comparison of correct solutions for the two rules which correspond to the normal thematic and abstract sentences is Thematic: 22%, Abstract: 6%. Thus in both these studies there is a low overall facilitatory effect of thematic materials and some difficulty in interpreting the significant factors found because they conflict. They also record a higher than usual rate of 'other' combinations (e.g. see Table 8b), and it is a pity that there is no closer examination of these, since a comparison of FC ($\overline{q}$) solution rates might have been profitable; and on a related point, it is regrettable that the analyses were of 'insight' scores based on the frequency of correct or partially correct combinations. We may now judge this to have been an error, though perhaps it was not at the time. Taken together, the findings of these two experiments and the caveats expressed above make it difficult to draw reliable conclusions from them, and one certainly cannot look to them for strong confirmation of the effect of thematic materials in the Selection task.

There have been other reports of correct solution rates under thematic materials that did not match up to expectations. Van Duyne (1976) reports a pilot study using "non-arbitrary and obviously commonplace sentences" in which "performance appeared to be surprisingly low". Presumably this was why it remained a pilot study. Lunzer, Harrison & Davey (1972) conducted a series of experiments using not only thematic materials, but a 'reduced' procedure in which only consequent items appeared on the selection cards and successive

presentations and explanations ('therapies') as well. Their thematic materials concerned pictures of and rules about lorries of different colours which could be laden or unladen. This is what they found. In their first experiment, there appeared to be a facilitation of correct responding under thematic materials, but on closer examination it appeared that this effect arose from an interaction with the 'reduced' presentation, and the experimenters remark that "both facilitating conditions are essential to produce (correct selections) with appreciable frequency." Unfortunately, this paper, like the ones above, concerns itself almost exclusively with $p\bar{q}$ and $pq\bar{q}$ selections, lumping all the 'others' together, so again it is not possible to assess the changes in $\bar{q}$ (FC) frequencies in the 'complete' (i.e. with all four selection cards) condition. Giving a second presentation of the problem after an explanation of it had little effect in the 'complete' form, and when the complete thematic problem was presented second without an explanation, no subjects at all produced the correct solution. In the second experiment, the second problem was standardised as a normal abstract task, to assess the extent of any transfer from prior exposure to different forms of the task. Only exposure to a prior abstract task seemed to lead to improved performance on the second problem; no such therapeutic value was found after a complete thematic task. This is not too surprising seeing that only one of the 16 subjects got the initial thematic problem right. The third experiment refined the technique of the second, using a group of graduate students as subjects instead of the sixth-form pupils used in the first two. Each subject was given three tasks in the following order: reduced thematic, complete abstract, explanation, complete abstract. Over a third of the subjects were completely correct on the first task, and a total of

42% were completely or partially correct on the first abstract task; there was no further improvement on the second abstract task, even though there was an intervening explanation. Again, any improvements in performance attributable to thematic materials arise only in the reduced presentation. All these results lead us to conclude that, in this study, any thematic materials effect was strictly interactive, manifesting itself only in conjunction with additional procedures having the same aim, both when performance on a thematic task and transfer to another task are considered.

These then are the studies which failed to produce the thematic effect they seemed to from a distance, and which urge caution in talking about a simple effect. However, there are several more studies which really do show dramatic increases in performance under thematic materials, and foremost among them is that of Johnson-Laird, Legrenzi, & Legrenzi (1972). Their thematic materials were envelopes, sealed or unsealed, bearing stamps of different values, and the rules were (e.g.) 'If a letter is sealed, then it has a 50 lire stamp on it'. They used 'only if' rules as well as 'if then', with no difference being found between them, and if results from these rules are pooled, we see that 81% of subjects selected the correct combination of envelopes in the thematic condition compared with 15% in the abstract condition (letters and numbers), a clear-cut result if ever there was one. However, if we pause for a moment and consider the precise nature of the task from the subjects' point of view, an alternative to the notion that the thematic presentation facilitates insight into the logical nature of the task by invoking a "sense of reality" presents itself. The subjects were asked to imagine that they were postal workers sorting letters, with a rule about the value of stamp which a sealed letter should carry.

Do they really need to be armed with insight into conditional logic to be able to pick out an understamped sealed letter? It seems at least as likely that people know perfectly well in advance, through common experience, that understamped letters are illicit, and simply act on this knowledge, i.e. they use a "detective set" (van Duyne, 1974). The fact that the item is also logically correct for this task is almost incidental. One could justly ask when a reasoning task stops being a reasoning task and becomes a memory task; thematic materials should surely be thematic enough to be interpreted as meaningful sentences, but not so thematic that they promote a correct solution by pointing straight at the correct items through other, non-reasoning, processes, otherwise the problem remains untouched.

An experiment by van Duyne (1976) could be regarded in a similar light. This experiment used a much more 'personal' procedure than is usual in this line of research: subjects composed their own conditional sentences, under instructions to avoid statements of equivalence, and the experimenter selected two of them for the task. He then tested the subjects by describing the antecedent and consequent items and their negations, and asking the subjects whether they would call for any additional information (a mental equivalent to turning the cards over) knowing each state of affairs indicated by the item to be true. Van Duyne thought he had found a difference between the treatment of sentences which subjects assumed to be always true and those assumed to be only sometimes true, but thanks to Pollard & Evans (in press) we now appreciate that this difference lay only in the explanations the subjects gave for their choices (van Duyne only scored a choice as correct if it was accompanied by an explanation expressing the appropriate logical turn of mind). If the actual choices

themselves are examined, the difference disappears. However, difference or no difference, it is still the case that in this experiment 93% of subjects chose $\bar{q}$ (FC), and of course all the sentences were thematic - none of van Duyne's subjects composed rules about letters and numbers or coloured shapes. Does not the act of sentence generation, though, when conditionals are involved, necessarily also presuppose a concurrent generation of falsifying instances? Surely, in deciding for himself that sentence is not an equivalence and is always or sometimes true, the subject must think up contingencies which could violate his rules? He will not utter them at the time because he is not asked to, but he is ready with them when, a few minutes later, he <u>is</u> asked, in a roundabout way, to produce them. This argument would also apply to the experiment by Pollard & Evans, who used the same procedure and found a difference in FC selections between rules considered to be true (both always and sometimes) and rules considered false. This interesting though perhaps unsurprising finding must also be tempered by the selection frequencies of the other items, which, except for the normally high TA frequency, were very much higher than usual.

Returning for the moment to the Johnson-Laird et al. study: it differs in another respect from the present one, in the instructions given to the subjects. In the former experiment, subjects were asked to indicate the envelope they would turn over "to discover whether or not they violate the rule", whereas the present experiments' instructions were to decide on the cards to be selected "to find out whether the rule has been obeyed or not" (Experiments 1 - 3) or "to decide whether the claim is true or false" (Experiments 4 and 5). The emphases in these two forms of instruction are different in two ways:

firstly, the former 'violate' instructions emphasise the cards themselves rather than the rule, and secondly they specifically raise the prospect of falsification by their use of the term 'violate the rule'. This may sound like a trivial point, but at least one researcher (Bracewell, 1974) has advocated the use of 'violate the rule' instructions as an essential prerequisite for subjects to embark on the task under the correct premises. 'Violate' instructions are also used by van Duyne (1974) in a study of different liguistic expressions of material implication with abstract and thematic materials. He used four types of sentence: 'Every p is q', 'If p then q', 'Not-p or q', and 'Not (p and not-q)'. Only in the cases of the first two did thematic materials facilitate correct selections. In addition to 'violate' instructions, van Duyne also directed his subjects to choose "those cards" to turn over, i.e. he specified that more than one card should be selected. This may have biased the subjects' behaviour, but such a bias would probably only have been slight, as 15% of all solutions were of p alone. The failure of thematics to facilitate correct solutions on the disjunctive form of the problem is confusing in the light of Wason & Johnson-Laird's (1969) finding that a disjunctive rule form in itself promoted correct selections, even in an abstract task. This result was not replicated in van Duyne's abstract condition.

The one study which has not been re-examined so far is the original thematic materials experiment by Wason & Shapiro (1971). That is because none of the above remarks can readily be applied to it - there are no obvious shortcomings in its design or analysis (apart from the excusable lack of reporting of independent FC frequencies). All that can be said about it by way of criticism is that the fifth

experiment in the present series was an exact repetition of it, but failed to repeat its results.

Having reviewed the previous published research on the comparison between abstract and thematic materials in the Selection task, and taking the results of the present experiments into account, the inescapable conclusion is that there is no evidence for a singular thematic materials effect of the type which has been assumed in the past. Lunzer's (1975) statement that "the difference between the familiar and abstract material is indisputable", and others in the same vein, are at best an oversimplification. That is not to assert that there is no such thing as an effect due to thematic materials at all, but simply to propose that for an effect to appear, there must be other contingencies in a Selection task experiment which conspire to the same end - the facilitation of correct solution. Among these, as we have seen, can be numbered long-term memory or experience, specific instructions, and special procedures such as 'reduced' arrays, individual discussion of items, and self-generated sentences. Thematic materials need to interact with such contingencies, or vice-versa, for facilitation of correct solution to occur.

In the next chapter, the idea of establishing solution sets in subjects through the manipulation of specific task contingencies is developed further. An experiment is reported in which a selection task is preceded by pretraining on a similar deductive task, treatment of this prior task being varied by instructions.

## CHAPTER 6

The object of the next experiment was to attempt to influence performance on a Selection task by establishing sets for different solutions. This arose from some of the points in the discussion of the published literature above, which leads to the suggestion that certain procedures in the period prior to undertaking a Selection task, as well as the conduct of the task itself, could affect performance. Particularly relevant here is the role of instructions, which it was pointed out could induce biases to falsify in some experiments but not in others, owing to different emphases within them. This kind of bias, should it exist, and should it play a significant role in task performance, might be even more potent if acquired through some kind of prior experience on a similar task. Wason & Shapiro (1971) attempted this by exposing subjects to 'therapies' - constructing and evaluating verifying and falsifying instances - before a Selection task. Little influence on Selection task performance was found. However, it might be possible, using instructions and prior experience, to induce an orientation to verify or falsify in the task. Accordingly, it was decided to give an ordinary abstract Selection task to two groups of subjects; each group would perform a truth-table task (see Chapter 3) beforehand, but one group would be instructed to judge which truth-table cases would make a conditional rule true, while the other would judge which instances would falsify the rule. Such a procedure should cast some light on the reality of verification and falsification biases in the Selection task, and would enable an assessment of the effectiveness of the truth-table task as a predictor of Selection task performance. Truth-table evaluations have been used before as 'therapies' in the Selection task (the relevant studies are summarised in Chapter 4), but not in an attempt at

differentially influencing performance. It was therefore expected that more $p\bar{q}$ combinations and $\bar{q}$ selections would occur when subjects were pretrained to falsify, rather than verify, a conditional rule. It was also decided to use response latencies in the truth-table tasks as a backup measure - perhaps if falsification is inherently more difficult than verification, as has been suggested, falsifiers should take longer to respond than verifiers.

## EXPERIMENT 6

### METHOD

#### Subjects

Thirty-two male and female students of Plymouth Polytechnic served as subjects. They were paid volunteers with no experience of these tasks, and were tested individually.

#### Task and Materials

(a) Truth-table tasks. There were two groups of subjects, both performed a set of 24 truth-table evaluation tasks, one having to judge which contingencies would verify a conditional rule and the other judging the falsifying contingencies. Six AA rules were presented, the four logical cases to each rule making up the 24 evaluations each subject had to make. Chapter 3 and Table 5 give more detailed information on the relation between conditional sentences and truth-tables. Four sets of rules were prepared, all concerning letter-number pairings, e.g. 'If the letter is a U then the number is a 4'. Twenty-four cards were prepared, each bearing one of the six rules and a separate truth-table case; they were presented one by one in a tachistoscope which was linked to a response switch and an electric timer. Rules were presented in random order, with all four instances, also in random order, after each rule.

(b) Selection task. Both groups performed an abstract Selection task with rules about letter-number pairs (as above) and cards with a letter on one side and a number on the other. Two alternative sets of rules and cards were made up, the subjects receiving one or other of these at random.

Procedure

Subjects were assigned to verifying and falsifying groups alternately. Treatment of both groups varied only in the instructions read out to them; after thanking the subject for coming, and familiarising him with the equipment, these were:

"This experiment is concerned with how people reason; please don't think of it as an intelligence test, because it isn't one.

You will be presented with a set of rules, and each rule will be presented with an instance which could mean that the rule is true or false in relation to that instance. You will be given 6 rules with 4 instances for each rule, making 24 presentations in all.

Your job will be to look at each instance for each rule, and decide whether that instance means that the rule is true (false). You indicate your decision by using the switch marked 'yes' and 'no' in front of you. If you decide that the instance you are presented with does mean that the rule is true (false), press it towards 'yes'; if it does not mean that the rule is true (false), press it towards 'no'.

The rules and their instances will be presented one by one in the T-scope. Here is an example of what you will be seeing on the screen. (Example card shown). At the end of the 24 trials you will be given another test, but only one rule will be involved in that".

Subjects were given a 'dry run' with the sample card in the

tachistoscope. They were asked if they had understood the instructions, which were repeated if doubts were expressed, and told that the timer, which was in full view (although the subjects could not see their times displayed), was for additional information only, that no time limit was in operation, and that they should come to their decisions in their own time. Each trial then consisted of placing a card in the tachistoscope, starting it, and thereby the timer, off, and recording the time between this and the subject's pressing of the yes/no switch. which stopped the timer and darkened the tachistoscope screen. The direction of the decision was also recorded. After the 24 truth-table trials, each subject was given a Selection task. For this, a rule was read out and also presented to the subject in written form, and the four cards placed on a table in random order. The subject was asked to indicate "which of the cards would need to be turned over to decide whether the rule is true or false". After recording the selections, a full debrief was given.

RESULTS

(i) Selection tasks. As the analysis of the Selection task data in this experiment is rather less involved than that for the prior truth-table tasks, it is dealt with first. Table 16 shows the frequencies of individual and combined card selections for both groups, and it will be immediately apparent that prior training on the truth-table task had no effect. The individual card selections were tested by the Fisher Exact test, and no difference between the verifying and falsifying groups was found. The large number of 'other' combinations in the falsifying group is most likely due to the occurrence of four $p\bar{p}q\bar{q}$ selections.

TABLE 16 (a) Number of times each card was selected in the Verifying and Falsifying groups in Experiment 6 N = 16 in each cell.

| Card | Group: Verifying | Group: Falsifying |
|---|---|---|
| TA | 14 | 15 |
| FA | 4 | 7 |
| TC | 7 | 10 |
| FC | 4 | 5 |

(b) Frequency of various selection combinations. N = 16 in each group.

| Combination | Group: Verifying | Group: Falsifying |
|---|---|---|
| TA | 3 | 3 |
| TA-TC | 6 | 6 |
| TA-TC-FC | 1 | 0 |
| TA-FC | 2 | 0 |
| Others | 4 | 7 |

(ii)Truth-table tasks. The truth-table paradigm was described in Chapter 3, but a short recap will be useful here. It will be recalled that there are four possible truth-table contingencies for any conditional rule, corresponding to the four combinations of affirmed (true) and denied (false) antecedent and consequent. Each of these contingencies may be judged to make the rule true or false, or to be irrelevant to it. In this experiment subjects were instructed to respond 'Yes' to items which made the rule true or false, depending which group they were in, so the 'irrelevant' category remained implicit. Table 17a shows the frequency with which each contingency received 'yes' responses for the verifying and falsifying conditions. In most cases, the double-affirming (TT) contingency is confirmed as the only definitely verifying case, although only 6 of the 16 subjects maintained the choice of TT alone over all 6 presentations of the task. Nine of the subjects also classified FF (the double-negating instance) as verifying at some time. In the falsifying group, responses are more variable: the TF case (true antecedent-false consequent) is most often classified as falsifying, but FT and FF are also similarly classified a high proportion of the time, and 13 of the 16 subjects classified a combination of all three cases as falsifying at some point.

The response latency measures were subjected to a 4 x 6 x 2 (truth-table case X blocks X groups) analysis of variance. Significant effects were found due to blocks, or successive presentations of the rules ($F\ 1, 30 = 44.92$, $p < .001$) and truth-table case ($F\ 1, 30 = 28.32$, $p < .001$). The full table is given in Appendix B. The first is a simple practice effect - subjects respond more quickly as the task progresses. The second shows the influence of verifying the different

TABLE 17 (a) Percentage frequencies of 'yes' responses to each truth-table case under each condition. There is a possible maximum of 96 responses in each cell (16 Ss. X 6 rules).

| | Truth-table cases | | | |
|---|---|---|---|---|
| Group | TT | TF | FT | FF |
| Verifying | 100 | 0 | 8 | 18 |
| Falsifying | 9 | 86 | 77 | 60 |

(b) Frequency of truth-table classification as combinations. N = 96 in each group

| Verifying group | % | Falsifying group | % |
|---|---|---|---|
| TT | 80 | TF-FT-FF | 52 |
| TT-FF | 11 | TF-FT | 22 |
| TT-FT-FF | 6 | TF-FF | 6 |
| TT-FT | 2 | TF | 7 |
| | | TT | 8 |
| | | FT | 2 |
| | | FT-FF | 1 |
| | | TT-FF | 1 |

components of the conditional; mean response times (in seconds) to the four truth-table cases were TT: 3.61, TF: 4.45, FT: 6.16, and FF: 6.77. The same effect and ordering was found by Evans & Newstead (1977); evidently verifying the antecedent uses up more time than does verifying the consequent, and the two effects summate. There was no effect of groups, so apparently neither task, verifying or falsifying, was harder than the other.

## DISCUSSION

The Selection task data have a familiar look, with no effect of pretraining for verification and falsification being found. This is perhaps not too surprising, as the history of 'therapies' in this paradigm is not an illustrious one. We may conclude from these results either that the task is immune to logical biases, or that the truth-table and Selection tasks were sufficiently distinct from one another that any bias generated by the former was dissipated in the latter. The first possibility is a good one; in the review of previous work we have seen how the evidence for rational biases (e.g. Wason's verification bias) has been called into question by later inquiries from a different standpoint (e.g. Evans' matching bias), and how the efforts to deflect subjects from the well-trodden tracks of the Selection task have been characterised by an almost unanimous fruitlessness. The second possibility also has merit however, and this can be seen in a closer examination of each subject's truth-table classifications and subsequent card selections. In the verifying group, 80% of the classifications were of TT alone as the verifying instance; for logical consistency this should lead to a majority of selections of the TA-TC (pq) combination, since both cards potentially bear the

TT case, and in fact the pq combination is the modal one, though not to the extent of an overall majority. All three subjects who selected the p card alone classified TT alone as the verifying case throughout, a reflection perhaps of the salience of the antecedent which other investigators have noted. Apart from these links between logical errors on the two tasks, there is also a slight suggestion of transfer of logical competence between them: of the three subjects who at some point classified a combination of TT, FT, and FF as verifying, two went on to select the TA-FC ($p\bar{q}$) combination of cards, and of course both this classification and selection are correct according to the formal logic of material implication.

These pieces of encouragement for the logicist do not carry over so well to the falsifying group. The most popular falsifying combination was TF-FT-FF, with 13 out of 16 subjects choosing it at some time during the task. This should lead to the turning over of all four cards, since all of them could have one or other of these contingencies on them, but only four subjects selected the $p\bar{p}q\bar{q}$ combination. Nine subjects selected p alone or pq, and eight of them had at some point classified FF as falsifying; neither the p nor the q card could carry an FF instance. In this group then the connection between behaviour on the truth-table tasks and performance on the Selection task seems to be only tenuously connected, and on this basis one must be cautious in treating the former as a predictor of the latter. From a logicist point of view, the truth-table tasks could lead us to suppose that the subjects in both groups (the groups' responses were almost mirror images of each other) were using the truth-table for conjuction -'p and q' The selection tasks do not confirm this: half the subjects do not choose either the potentially

verifying combination (pq) or the falsifying combination ($p\bar{p}q\bar{q}$) for a 'p and q' truth-table. That the Selection and truth-table tasks should be regarded as psychologically distinct paradigms is not a new conclusion (see, e.g. Evans, 1978), and Evans' work also leads us to an alternative view of the data from this experiment: perhaps the subjects in both tasks were not using truth-tables at all, but responding according to the psychological processes of interpretation and operation - they were engaged largely in matching behaviour. If so, the matching going on in the truth-table tasks must have been of a rather crude nature, for it seems that the effect must have been for all contingencies which mismatched the items named in the rules in some way to be seen as falsifying, or at least non-verifying. This kind of pattern was observed by Paris (1973) working with children; he only used the T and F categories. The higher than usual proportions of FT and FF cases classified as falsifying bear this out: in Evans' 1972b experiment, where subjects had to construct instances to verify and falsify rules, only 27% and 33% constructed FT and FF cases, respectively, to falsify, and in a later study using an evaluation procedure, 46% classified FT and only 4% classified FF as Falsifying (Evans, 1975). Both these studies used an 'irrelevant' category, implicit in the first and explicit in the second, whereas in the present experiment the procedure seems to have pressed the subjects into using a bivalent truth function. They did not perceive, or use, an 'irrelevant' category.

In this experiment the interest has shifted from the Selection task to the truth-table task, and in the context of the first five experiments some interesting possibilities have opened up. The original intention was to extend the pretraining format adopted here to thematic as well as abstract materials, to explore the nature of

their interactive effect in the Selection task. This could still be done sometime, but at the moment the truth-table task seems to offer prospects at least as fruitful to similar ends. It is an under-used paradigm in deductive reasoning research, most probably because it has been used more as an adjunct to inference or Selection tasks than as a problem in its own right. It has some particular advantages over the Selection task: only one experiment using thematic materials in the truth-table task has been done (Rips & Marcus, 1977), which is quite surprising considering it is a paradigm said to reflect 'natural' logical ability; it is a simple task from the subject's viewpoint, and therefore less susceptible to the kinds of intervening variables explored in Experiments 1 - 5; and it is more flexible than the Selection task in its adaptive capacity in the laboratory. This has been amply demonstrated by Evans & Newstead (1977), who used a technique of splitting response times (pioneered by Trabasso, Rollins & Shaughnessy, 1971) to separate interpretative and operational factors and investigate certain linguistic variables (see Chapter 3). There is an obvious extension of this work to include thematic materials - the effects of interpretation and operation might be different for them - and for this investigation to shed some light on what was happening in the first five experiments.

## CHAPTER 7

In the next experiment we extend the comparison between abstract and thematic materials, begun in Experiments 1 - 5, to the truth-table task used in the pretraining part of Experiment 6. An evaluation procedure is used, with an 'irrelevant' category added, and with systematically negated rules as well as the ordinary 'If p then q'. There is a single reason behind both these adaptations. In the last experiment, we saw that subjects' judgements of the truth-table cases were open to two interpretations: they could be regarding the double-affirming (TT) case as the only verifying instance and all the others as non-verifying or falsifying, or they could be operating on the basis of a crude matching bias, whereby the instance which simply matched both rule components - in this experiment also the TT case - was verified, with 'false' being treated as synonymous with 'mismatching'. Evans (1972b, 1975) had found that the effect of matching bias in the truth-table task was for mismatching items to be considered irrelevant, i.e. neither verifying nor falsifying, to the rule, but in Experiment 6 the forced-choice task seemed to preclude the use of an 'irrelevant' category, subjects including this at least partially in their 'false' responses. In the next experiment the third category is explicit, which should clarify the effects of matching.

The use of negated rule components will allow the separation of matching from logical biases. It will be recalled from the Review section that these are confounded in the normal AA rule, where the verifying and falsifying items are also the matching and mismatching ones, respectively. Negatives in the rules disrupt this coincidence, and by varying the presence of negation systematically in the antecedent and consequent the separation of matching value and logical

case can be balanced exactly; this is shown in Table 6. Using this procedure, Evans (1972, 1975) found that the ruling out of instances as irrelevant, and to a lesser extent the classification of relevant items, was determined largely by matching value and not the logical status of items. Similar results were obtained by Evans & Newstead (1977). All these studies were limited to abstract materials, and it is of obvious interest to test whether their findings will generalise to a thematic presentation. Previous research on the Selection task might assure us they would not, but the first five experiments here have shown that we must apply such predictions with caution, if we apply them at all. However, it was discovered in the re-examination of the literature that an effect due to materials was likely to arise in the presence of intervening variables, and one such may be a simplified task: the truth-table task is simpler for the subjects, being both easier to instruct and presenting less of a problem in only requiring classification. It is therefore quite difficult to predict just what the response frequencies will be on the thematic form of the task, though the Evans studies (Chapter 3) allow us to expect with a fair degree of precision what the Abstract results will look like.

The Evans & Newstead experiment used a procedural innovation - split response times - which enabled them to make a distinction between competing interpretations of their findings which was not possible by recourse to the response frequencies alone. Their technique was to split the time taken to understand the rule from the time taken to make a truth-table evaluation when an instance was presented. This is potentially a valuable measure where content differences are involved, since again we might be confronted with similar response frequencies, but there might be interpretational or operational differences between abstract and thematic materials, reflected in latencies, which the

behavioural indices were not sensitive enough to pick up. For instance, it may be that thematic rules are easier to comprehend than abstract rules, but that the reasoning operations on them are equally difficult; thus the eventual decisions for both types of content could look quite similar, but there could be underlying effects revealed by the latencies (in this case comprehension times) which would alert us to the possibility of different processes taking place.

## EXPERIMENT 7

### METHOD

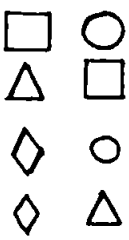
#### Subjects

Twenty-four male and female students of Plymouth Polytechnic served as subjects. As usual, they were paid volunteers with no experience of this type of task; they were tested individually.

#### Task and Materials

The truth-table evaluation task was used. This involves presenting a subject with a conditional rule followed by one of the four truth-table contingencies (see Tables 5 and 6), and asking him to say whether the instance verifies the rule, falsifies it, or is irrelevant to it. In this experiment the four systematically negated conditionals were used, allowing the balanced separation of matching value and truth-table case which is set out in full in Table 6. Each subject received the four rules in random order; with the four instances following each rule, also in random order. Two sorts of materials were used: outline shapes (abstract) and foods and drinks (thematic; see Expts. 1 - 4). Two examples of these rules are given in Table 28 as an illustration,

TABLE 18 Two examples of the kinds of rule used in Experiment 7, with the logical and matching status of the four contingencies for each. For a fuller illustration of these, and an explanation of the notation, consult Tables 5 and 6.

| Abstract rule (NA) | | Truth-table case | Matching value |
|---|---|---|---|
| Every time there is not a diamond on the left, there is a circle on the right. |  | TT | $\bar{p}q$ |
| | | TF | $\overline{pq}$ |
| | | FT | $pq$ |
| | | FF | $p\bar{q}$ |

| Thematic rule (AN) | | | |
|---|---|---|---|
| Every time I eat chicken I do not drink brandy | Chicken & Whisky | TT | $\bar{p}q$ |
| | Chicken & Brandy | TF | $pq$ |
| | Pork & Gin | FT | $\overline{pq}$ |
| | Fish & Brandy | FF | $\bar{p}q$ |

along with the matching and logical values of their four instances. In the experiment, each rule and each instance were typed or drawn on separate tachistoscope cards.

Procedure

The cards were fed into a two-field tachistoscope which was connected to a start key, a decision switch, and two electric timers. The progress of a single trial is described in the instructions read out to the subjects. Each subject was allocated alternately to the Abstract and Thematic groups and familarised with the tachistoscope; as the instructions vary only slightly between the two groups, both forms are set out together:

"The basic setup is this: on cards you will see four rules, one at a time.
The rules define

the position of different kinds of shapes on a card. (for the Abstract group)

the foods and drinks I have together. (for the Thematic group)

Each rule will be followed by four instance cards, again one at a time, and the cards will show

different combinations of shapes. (Abstract)

different combinations of foods and drinks. (Thematic)

Your job is to decide whether the instance you are presented with conforms to the rule, contradicts it, or is irrelevant to it. You control the presentation of materials and record your decisions by using the two switches in front of you. This is how they work.

You use the morse key on the left to call up the test cards. On the first press, the rule will appear; when you have understood the rule, press again and the instance will appear. When the instance

appears you have to decide whether it conforms to the rule, contradicts it, or is irrelevant to it, and you record your decision using the two-way switch marked 'true' and 'false' on your right. Press it towards 'true' if the instance conforms to the rule, towards 'false' if it contradicts the rule, and backwards and forwards if it is irrelevant. You will have to call up each rule four times, once for each instance, as the lights go out in the tachistoscope after each response".

The subject was then asked if he had understood the instructions (repetition was given if doubts were expressed) and given a practice run using an AA rule and a TT instance which did not figure in the experiment - he was not given any feedback about the merits of his practice response. In summary, the sequence of events per trial was: tachistoscope loaded with rule and instance cards; subject signalled to start; start key pressed once; start key pressed again, comprehension time (CT) recorded; decision switch pressed, verification time (VT) and direction of decision (indicated by lights at the rear of the machine) recorded; instance card or instance and rule cards (after every fourth trial) changed. The timer displays and decision lights were out of the subjects' view, although if a subject asked if he was being timed (several did so), he was told that he was, but that there was no time limit operating. A full debrief was given at the end of the test session.

RESULTS

(i) Response frequencies

Table 19a gives the total response frequencies for 'true', 'false', and 'irrelevant' responses to all the logical cases on all

TABLE 19 (a) Frequencies of 'true' (T), 'false' (F), and 'irrelevant' (?) responses to the four truth-table cases on the four rules for the Abstract and Thematic groups. N = 12 in each cell. Data from Experiment 7.

| | | Abstract | | | | Thematic | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Truth-table cases | | | | | | | |
| Rule | | TT | TF | FT | FF | TT | TF | FT | FF |
| | T | 12 | 0 | 1 | 2 | 12 | 0 | 2 | 2 |
| AA | F | 0 | 10 | 6 | 4 | 0 | 10 | 6 | 0 |
| | ? | 0 | 2 | 5 | 6 | 0 | 2 | 4 | 10 |
| | T | 11 | 0 | 5 | 4 | 10 | 2 | 1 | 5 |
| AN | F | 1 | 12 | 1 | 2 | 1 | 10 | 0 | 0 |
| | ? | 0 | 0 | 6 | 6 | 1 | 0 | 11 | 7 |
| | T | 11 | 3 | 2 | 1 | 10 | 0 | 2 | 8 |
| NA | F | 0 | 6 | 9 | 4 | 0 | 2 | 9 | 2 |
| | ? | 1 | 3 | 1 | 7 | 2 | 10 | 1 | 2 |
| | T | 10 | 2 | 5 | 7 | 4 | 4 | 8 | 5 |
| NN | F | 0 | 7 | 2 | 1 | 0 | 5 | 2 | 2 |
| | ? | 2 | 3 | 5 | 4 | 8 | 3 | 2 | 5 |

(b) Frequency of 'irrelevant' responses to the four truth-table cases pooled across rules as a function of matching value. Data from Experiment 7. N = 12 in each cell.

| | Abstract | | | | | Thematic | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Matching value | TT | TF | FT | FF | Total | TT | TF | FT | FF | Total |
| pq | 0 | 0 | 1 | 4 | 5 | 0 | 0 | 1 | 5 | 6 |
| $p\bar{q}$ | 0 | 2 | 5 | 7 | 14 | 1 | 2 | 2 | 2 | 7 |
| $\bar{p}q$ | 1 | 3 | 5 | 6 | 15 | 2 | 3 | 4 | 7 | 16 |
| $\bar{p}\bar{q}$ | 2 | 3 | 6 | 6 | 17 | 8 | 10 | 11 | 10 | 39 |

the four rules. On inspection, it seems that in both groups there is an overall tendency to give the TT case as verifying and to a lesser extent the TF case as falsifying, which is in line with the findings of Evans (1972, 1975). The lower incidence of these correct logical choices on the NA and NN rules, found by Evans, is also reproduced here, as is the greater variability in the classification of the FT and FF cases. There does, however, seem to be a difference on both these points between the groups, and not in the way one might have expected: the classification of TF as falsifying on the NA rule almost disappears in the Thematic group, and the classification of the FT and FF cases as irrelevant on the AN and AA rules is almost unanimous in the Thematic group but not in the Abstract group. In all three cases the item involved is the double-mismatching ($\overline{pq}$) contingency.

The effects of matching bias were tested both within and between the two groups. As we have seen before, the matching effect in the truth-table task consists in a tendency to call mismatching items irrelevant (where only 'true' and 'false' responses are allowed, Experiment 6 would lead us to expect more 'false' responses to these items). Accordingly, it is possible to test for the effect of matching on both antecedent and consequent: there should be more 'irrelevant' judgements, if matching is exerting an influence, of the $\bar{p}q$ and $\overline{pq}$ items compared with the pq and $p\bar{q}$ items (the antecedent effect), and more 'irrelevants' to the $p\bar{q}$ and $\overline{pq}$ items compared with the pq and $\bar{p}q$ items (the consequent effect). One-tailed sign tests on these responses from individual subjects were carried out to test these comparisons in both groups: all were significant ($p < .05$) in the predicted directions. Thus both Abstract and Thematic groups were significantly influenced by matching in their evaluations.

What of the comparisons between groups, to test the observations made above? This entailed a comparison of the relative sizes of the matching effects of the groups, and the Mann-Whitney U test was used for this, corrected for ties. Each matching contingency will have a possible maximum of four 'irrelevant', or other, judgements of it, as each is represented once in the four rules, and as the comparisons for the antecedent and consequent matching effects are between pairs of instances, there is a possible range of scores for each subject between +8, where all mismatching items are judged irrelevant and all matching items are not, and -8 where the reverse is the case. These scores allow an index of the sizes of the matching effects. These tests were two-tailed, as no firm a priori predictions of group differences were made. The comparisons of the antecedent and consequent matching effects yielded the same result: both were significantly greater in the Thematic group. (Antecedent effect: $U = 30.5$, $p < .02$; Consequent effect, $U = 29$, $p < .02$). Looking at Table 19b, one can see that this finding has its source primarily in the responses to the $\overline{pq}$ instance alone - this item was ruled out as irrelevant 81% of the time in the Thematic group and only 35% of the time in the Abstract group.

(ii) Response latencies

The comprehension times (CTs) and verification times (VTs) were submitted to a log. transformation and analysed separately. CTs were subjected to a 2 x 2 x 2 (Groups x Polarity of antecedent x Polarity of consequent) analysis of variance, with repeated measures on the last two factors. Significant main effects due to all three factors were found: Groups ($F = 14.02$, $p < .01$), Antecedent ($F = 10.24$, $p < .01$), and Consequent ($F = 17.92$, $p < .001$), with a significant

interaction between Groups and Antecedent ($F = 9.35$, $p < .01$; all ratios were assessed on conservative degrees of freedom; see Appendix C). The mean CTs for all four rules in both groups are shown in Table 20: negating the antecedent and consequent seems to slow understanding under both types of materials, but it seems that an affirmative antecedent speeds up comprehension of thematic rules relative to all the other sentences.

Verification times (VTs) were first subjected to a 2 x 4 x 2 x 2 (Groups x truth-table case x Antecedent x Consequent) analysis of variance. The result was a four-way interaction, two lower-order interactions, and three main effects (Antecedent, Consequent, and Truth-table case; see Appendix D). The response frequencies however have shown that it is matching which seems to exert the greater influence over subjects' evaluations rather than truth-table case, and so the VTs were reanalysed with the three within - group factors replaced by a Rule factor and a Matching case factor. The analysis of variance was thus a three-factor, 2 (groups) x 4 x 4 one; see Appendix E. Using conservative degrees of freedom, two significant main effects and two interactions were found: Rules ($F = 14.29$, $p < .01$) and Matching value ($F = 9.47$, $p < .01$), and Matching x Rules ($F = 5.59$, $p < .05$) and Matching x Groups ($F = 5.09$, $p < .01$). The analysis by matching values is therefore justified both on the grounds of parsimony and the relation to the effects observed in the response frequencies, and it is this analysis which will be discussed from now on. It may be noted straight away that there is no overall difference in VTs due to the two types of materials, rather it is the pattern of latencies within the groups which differs. This may be confirmed by inspecting Table 21a, where the relevant mean latencies for inter-

TABLE 20 Comprehension times for the four rules (in seconds) in the two groups in Experiment 7.

| | Rules | | | | |
|---|---|---|---|---|---|
| Group | AA | AN | NA | NN | Mean |
| Abstract | 7.33 | 8.02 | 7.21 | 8.87 | 7.85 |
| Thematic | 3.58 | 4.55 | 5.71 | 6.99 | 5.21 |
| Mean | 5.46 | 6.29 | 6.46 | 7.93 | 6.53 |

TABLE 21 (a) Verification times in Experiment 7. Times (in seconds) for both groups ordered in terms of matching value.

| | | Matching value | | | | |
|---|---|---|---|---|---|---|
| Groups | | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ | Mean |
| Abstract | | 6.45 | 5.85 | 6.62 | 5.66 | 6.15 |
| Thematic | | 5.39 | 6.88 | 7.39 | 5.01 | 6.16 |
| | Mean | 5.92 | 6.37 | 7.01 | 5.34 | 6.16 |

(b) Verification times in Experiment 7. Times for each rule ordered in terms of matching value.

| | | Matching value | | | | |
|---|---|---|---|---|---|---|
| Rule | | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ | Mean |
| AA | | 2.86 | 3.83 | 6.86 | 4.23 | 4.45 |
| AN | | 4.31 | 6.11 | 7.91 | 5.48 | 5.95 |
| NA | | 5.62 | 6.73 | 5.43 | 6.09 | 5.97 |
| NN | | 10.89 | 8.80 | 7.82 | 5.54 | 8.26 |
| | Mean | 5.92 | 6.37 | 7.01 | 5.34 | 6.16 |

preting the groups x matching interaction are displayed. There is little difference between the contingencies in the Abstract group, and Scheffe' tests confirm this, but similar tests show that in the Thematic group the $p\bar{q}$ and $\bar{p}q$ items take longer to evaluate than do the pq and $\bar{p}\bar{q}$ items. As regards the matching x rules interaction, it seems that in general the order of latencies (assessed with the help of Scheffe tests) is AA < AN = NA < NN, which is the order one would expect on the basis of previous research (cf. Evans & Newstead, 1977, where a similar though non-significant ordering was found); the difficulty of assessing each individual item seems to vary between the rules.

## DISCUSSION

The first and most striking aspect of these data is the complete lack of evidence that truth-table classifications are any more closely allied to logic when rules are thematic than when they are abstract. Were it not for the results of Experiments 1 - 5 this would be an astonishing finding, but as it is the present data constitute both an important confirmation and an extension of the Selection task results. Not only has the lack of facilitation by thematic materials generalised to a distinct paradigm, there is also the rather suprising observation that in some circumstances the Thematic group were even less logical than the Abstract group - there was a nearly unanimous ruling out of the double-mismatching instance as irrelevant by Thematic subjects, whatever its logical consequences. This is echoed, somewhat paradoxically, in the analysis of verification times. We have become used to doubly-negated, or denied, or mismatched, sentences and instances bringing about extra difficulty, yet in Table 21a we see that the average VT for the $\bar{p}\bar{q}$ instance under thematic rules is

in fact the shortest recorded, so for some reason the judgement of the double-mismatching case seems particularly easy, especially when materials are realistic. (We defer speculation as to just what this reason might be to the discussion following the final experiment). The one area where thematic materials seem to have a positively beneficial effect is in comprehension times. These are generally shorter for the thematic rules, and the interaction observed between polarity and materials shows that thematic rules with affirmative antecedents were particularly easy, or at least particularly quick, to understand. This accords well with prior intuition and subjects' comments.

It would be premature to pursue such detailed discussion, backed as it is mostly by the inspection of mean solution times from just 24 subjects. There are also some differences between the latency results of this experiment and the only comparable one, that of Evans & Newstead (1977). They found that effects observed in the comprehension times tended to carry over largely unchanged to the verification times, whereas this tendency is far less pronounced here (one must compare only the Abstract group's results). One possible reason for this difference could be in the manner of recording the VTs.. In Evans & Newstead (1977) when the subject pressed his key for the second time to call up the instance, the rule remained in view, but in the present study the instance replaced the rule. The latter procedure would seem to constitute a purer measure of both CT and VT: in the Evans & Newstead experiment the subject will realise after one or two trials that he does not need to be too sure about his understanding of the rule, since he can always review it when the instance comes up, but in the present experiment he will equally quickly appreciate that he must fully comprehend the rule before

proceeding to his examination of the instance. An additional slight divergence between the two experiments is in Evans & Newstead's observation of a simple main effect of truth-table case on VT, while in Experiment 7 the truth-table case effect was subsumed under a complex interaction and therefore practically uninterpretable. Evans & Newstead did not use the VT analysis by matching case which was found to be more useful here.

For these reasons then, a replication of the experiment is called for. There are also some questions raised here which may be answered by a more thorough investigation: is the extra-matching effect observed on thematic materials a function of the linguistic form of the rules used? An alternative conditional expression would settle this. Is matching of the kind we have seen limited to conditionals? The use of other sentences should give some indication; previous research with disjunctive rules in the Selection task (e.g. van Duyne, 1974; Wason & Johnson-Laird 1969) suggests that matching bias may not generalise to this kind of rule at least. The use of a rule-form which is immune to matching may help to elucidate the differences between the materials which have arisen here in a truth-table task but not in several Selection tasks. The next experiment therefore uses a more thorough procedure to extend the present investigation. Further discussion and speculation must await its outcome.

## EXPERIMENT 8

For the final experiment we examine further the effects found in Experiment 7. It is rather difficult to base firm conclusions on the latency measures taken in the previous experiment, as some of

the effects observed in the VTs are a little intricate and the data are drawn from a small population. The CTs seem clearer: thematic rules are apparently easier to comprehend, especially when the antecedent is affirmative. We would expect this to recur in a replication, but the VTs need further work before meaningful interpretation will be possible. The two additional rule-forms to be used in Experiment 8 should clarify the roles of materials in comprehension and evaluation difficulty: one of these will be the 'only if' conditional, the other the 'either or' disjunctive. The former has been found to share many of the behavioural characteristics of the 'if then' sentence but to have a slightly different meaning: it is affected by matching bias and is directional like the 'if then' form, but with its weight on the consequent rather than on the antecedent (Evans, 1977a, Evans & Newstead, 1977). Evans & Newstead found that comprehension was affected by the temporal order of the constituent items, such that the 'if then' sentence most comfortably expressed a relation where the antecedent preceded the consequent in time, the 'only if' form expressing the reverse relation. This factor is circumvented in Experiment 8 because neither the abstract nor the thematic materials carry such temporal connotations. The disjunctive is, of course, non-directional - it is a rule of alternation not condition. It has been found in the past to incur particular difficulty when its components are negated, but this conclusion comes mostly from inference studies (e.g. Johnson-Laird & Tridgell, 1972; Roberge, 1978), and the use of thematic materials has not been systematic. A disjunctive, thematic, truth-table task has never been reported, indeed the disjunctive has hardly been touched in truth-table experiments since

Johnson-Laird & Tagart (1969). The exploration of comprehension and verification of thematic and abstract disjunctives is largely virgin territory, and therefore of great interest.

Of more pressing concern than the latency analysis, which is always one step removed from a direct assessment of difficulty, is the pattern of evaluations found in the previous experiment. The two main findings were firstly that thematic materials did not lead to an improved logical performance, and secondly that in some circumstances this performance was actually worse under realistic rules, owing to the greater matching effects found, this in turn arising from a dominant regard of the double-mismatching case as irrelevant by Thematic subjects. The use of the 'only if' conditional rule-form should clarify whether this has something to do with the directionality of the 'if then' sentence or not - its slightly different apparent meaning might interact with thematic materials to affect its known susceptibility to matching. Responses to the disjunctive, which has been found to be immune to matching, should tell us whether extra-matching under thematic materials is a reaction peculiar to conditional reasoning or perhaps a reflection of some general strategy. Finally, the 'if then' rules should provide some much-needed replication, or otherwise, of the rather surprising findings of the previous experiment.

## METHOD

### Subjects

Thirty-two male and female students of Plymouth Polytechnic, recruited as paid volunteers and with no experience of tasks of this type, served as subjects. They were tested individually.

Task and Materials

Rules and instances were prepared on tachistoscope cards, as in Experiment 7. Because of the inclusion of an extra conditional, the wording of the first sentences was changed from 'Every time...' to 'If...then...'. The content of the abstract rules was also changed, from shapes to letters and numbers. This was because it would have been impossible, using the few shape words available, to construct enough rules and instances without repetition. The thematic materials were unchanged. Three rule-forms were used: 'If p then q' (IT), 'p only if q' (OI), and 'Either p or q' (EO), with the four systematically negated rules under the two types of materials being composed for each. No combination of antecedent and consequent occurred more than once in the rules or instances. Examples of the kinds of rules used appear in Table 22. With three rule-forms, four rules to each form, and four instances to each rule, each subject had 48 evaluations to make.

Procedure

Subjects were again allocated to Abstract and Thematic groups alternately. The equipment and the progress of briefing, trial and debriefing were the same as in the previous experiment, with the ordering of trials by a similar partial randomisation procedure: the order of presentation of rule-forms was randomised, as was the order of rules and instances, but all the rules for any one form and all the instances for any one rule were presented sequentially in a block. The wording of the instructions was modified to accommodate the new conditions but not altered substantially, so the instructions need not be reproduced in full again. The subjects were told that they would see 12 statements which would be in three forms and so would not be all the same; the statements would define "which letters

TABLE 22 Some examples of the kinds of rule used in Experiment 8

(i) 'If then' (IT) form.

If there is a J on the left, then there is a 7 on the right. (AA, Abstract)

If I do not eat mutton then I drink sherry. (NA, Thematic)

(ii) 'Only if' (OI) form.

There is a D on the left only if there is not a 4 on the right. (AN, Abstract)

I eat cheese only if I drink beer. (AA, Thematic)

(iii) 'Either or' (EO) form.

Either there is not a B on the left or there is a 9 on the right. (NA, Abstract)

Either I do not eat fish or I do not drink whisky. (NN, Thematic)

and numbers appear together as pairs" or "which foods and drinks go together in a set of imaginary meals".

RESULTS

(i) Response frequencies

Table 23 gives the frequencies of 'true', 'false', and 'irrelevant' responses to all the contingencies, and Table 24 the 'irrelevant' responses to the matching cases. The rule-forms will be considered one by one. Firstly, the IT form, which constitutes the replication of Experiment 7. The same trends emerge on inspection of Table 23 as on inspection of Table 19: an overall suggestion of TT given as a verifying case and TF as falsifying in both groups, and again a sharp difference between the groups in classifying these cases on the NA and NN rules, where they form the $\overline{pq}$ instance. This latter trend is confirmed in Table 24, where once again there is the striking increase in $\overline{pq}$ 'irrelevant' responses in the Thematic group. The same tests for the antecedent and consequent matching effects were performed as in Experiment 7, i.e. one-tailed sign tests, and again both comparisons were significant for both groups ($p < .01$, all tests). The Mann-Whitney tests for the sizes of the matching effects between the groups were both significant also: the antecedent and consequent matching effects were again larger in the Thematic group, (Antecedent: $U = 71.5$, $p < .03$; Consequent: $U = 44.5$, $p < .001$; one-tailed tests).

The same analysis was performed on the frequencies for the OI rule-form, and the frequencies of all three responses to the logical cases and the 'irrelevant' responses to the matching cases may also be inspected in Tables 23 and 24. Similar trends are apparent, and similar results arise from the analyses of the matching effects, which were the same as for the IT rules. Sign tests of the antecedent

TABLE 23 Frequencies of 'true', 'false', and 'irrelevant' responses to the four truth-table cases on each rule on the three rule-forms. N = 16 in each cell. Notation as in Table 19. Data from Experiment 8.

| 'IF THEN' | | ABSTRACT | | | | THEMATIC | | | |
|---|---|---|---|---|---|---|---|---|---|
| Rules | | TT | TF | FT | FF | TT | TF | FT | FF |
| | T | 16 | 0 | 3 | 3 | 16 | 1 | 1 | 0 |
| AA | F | 0 | 16 | 10 | 3 | 0 | 15 | 8 | 3 |
| | ? | 0 | 0 | 3 | 10 | 0 | 0 | 7 | 13 |
| | T | 15 | 0 | 4 | 7 | 12 | 1 | 1 | 11 |
| AN | F | 1 | 16 | 1 | 2 | 2 | 15 | 1 | 11 |
| | ? | 0 | 0 | 11 | 7 | 2 | 0 | 14 | 4 |
| | T | 15 | 1 | 2 | 9 | 14 | 0 | 1 | 9 |
| NA | F | 1 | 8 | 10 | 1 | 1 | 4 | 14 | 2 |
| | ? | 0 | 7 | 4 | 6 | 1 | 12 | 1 | 5 |
| | T | 13 | 2 | 5 | 11 | 2 | 2 | 4 | 12 |
| NN | F | 0 | 12 | 7 | 1 | 1 | 12 | 9 | 3 |
| | ? | 3 | 2 | 4 | 4 | 13 | 2 | 3 | 1 |
| 'ONLY IF' | | | | | | | | | |
| | T | 16 | 0 | 2 | 6 | 16 | 2 | 4 | 0 |
| AA | F | 0 | 15 | 11 | 1 | 0 | 13 | 6 | 1 |
| | ? | 0 | 1 | 3 | 9 | 0 | 1 | 6 | 15 |
| | T | 16 | 1 | 1 | 12 | 14 | 2 | 2 | 10 |
| AN | F | 0 | 15 | 4 | 1 | 1 | 14 | 3 | 4 |
| | ? | 0 | 0 | 11 | 3 | 1 | 0 | 11 | 2 |
| | T | 16 | 3 | 1 | 10 | 12 | 1 | 3 | 12 |
| NA | F | 0 | 6 | 14 | 2 | 3 | 7 | 13 | 4 |
| | ? | 0 | 7 | 1 | 4 | 1 | 8 | 0 | 0 |
| | T | 12 | 2 | 2 | 9 | 4 | 7 | 4 | 12 |
| NN | F | 2 | 12 | 12 | 5 | 1 | 6 | 9 | 4 |
| | ? | 2 | 2 | 2 | 2 | 11 | 3 | 3 | 0 |
| 'EITHER OR' | | | | | | | | | |
| | T | 3 | 16 | 16 | 0 | 2 | 12 | 11 | 0 |
| AA | F | 9 | 0 | 0 | 11 | 13 | 3 | 4 | 6 |
| | ? | 4 | 0 | 0 | 5 | 1 | 1 | 1 | 10 |
| | T | 11 | 7 | 11 | 2 | 10 | 5 | 3 | 9 |
| AN | F | 4 | 7 | 3 | 13 | 6 | 11 | 2 | 6 |
| | ? | 1 | 2 | 2 | 1 | 0 | 0 | 11 | 1 |
| | T | 13 | 9 | 5 | 3 | 12 | 3 | 2 | 9 |
| NA | F | 2 | 4 | 8 | 12 | 4 | 3 | 14 | 7 |
| | ? | 1 | 3 | 3 | 1 | 0 | 10 | 0 | 0 |
| | T | 3 | 14 | 12 | 1 | 2 | 13 | 13 | 3 |
| NN | F | 9 | 2 | 4 | 13 | 2 | 3 | 1 | 13 |
| | ? | 4 | 0 | 0 | 2 | 12 | 0 | 2 | 0 |

and consequent matching effects were significant for both groups ($p < .01$, all tests). The between-group Mann-Whitney tests were also significant (Antecedent: $U = 66$, $p < .01$; Consequent: $U = 67.5$, $p < .03$; one-tailed tests): both effects were greater in the Thematic group.

The picture for the EO rule-form is rather different. In the Abstract group, no matching effects were expected and none were observed; instead there was a significant tendency to rule out double mismatching and matching items as irrelevant relative to singly matching/mismatching items. ($p < .01$, two-tailed test). There were fewer 'irrelevant' classifications overall than on the conditional rule-forms, as one might expect from rules of alternation (cf. Johnson-Laird & Tagart, 1969). The 'irrelevant' response profile in the Thematic group is completely different (see Table 24) - there is now no trace of the symmetrical effect seen in the Abstract group, but a huge proportion of 'irrelevant' responses to the $\overline{pq}$ case. Tests for matching and the single v. double effect were done, but their value is questionable and their outcomes entirely predictable: there was significant evidence for both in the Thematic group. One does not need statistics to perceive the size and the source of the response frequency differences between the groups on the EO rule-form.

(ii) Response latencies

All latencies were again given a log. transformation. Comprehension times were subjected to a 2 x 3 x 2 x 2 (Groups x Rule-forms x Antecendent x Consequent) analysis of variance, with repeated measures on the last three factors (see Appendix F). Significant main effects due to Rule-forms ($F = 15.51$, $p < .001$), Antecedent ($F = 72.05$, $p < .001$) and Consequent ($F = 35.89$, $p < .001$) were found, and there were significant interactions between Antecedent and

TABLE 24 Frequency of 'irrelevant' responses to the four truth-table cases pooled across rules as a function of matching value; data from Experiment 8. N = 16 in each cell.

'IF THEN'

| Matching value | Abstract Truth-table cases | | | | | Thematic Truth-table cases | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TT | TF | FT | FF | Total | TT | TF | FT | FF | Total |
| pq | 0 | 0 | 4 | 4 | 8 | 0 | 0 | 1 | 1 | 2 |
| $p\bar{q}$ | 0 | 0 | 4 | 6 | 10 | 2 | 0 | 3 | 5 | 10 |
| $\bar{p}q$ | 0 | 2 | 3 | 7 | 12 | 1 | 2 | 7 | 4 | 14 |
| $\bar{p}\bar{q}$ | 3 | 7 | 11 | 10 | 31 | 13 | 12 | 14 | 13 | 52 |

'ONLY IF'

| | TT | TF | FT | FF | Total | TT | TF | FT | FF | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| pq | 0 | 0 | 1 | 2 | 3 | 0 | 0 | 0 | 0 | 0 |
| $p\bar{q}$ | 0 | 1 | 2 | 4 | 7 | 1 | 1 | 3 | 0 | 5 |
| $\bar{p}q$ | 0 | 2 | 3 | 3 | 8 | 1 | 3 | 6 | 2 | 12 |
| $\bar{p}\bar{q}$ | 2 | 7 | 11 | 9 | 29 | 11 | 8 | 11 | 15 | 45 |

'EITHER OR'

| | TT | TF | FT | FF | Total | TT | TF | FT | FF | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| pq | 4 | 2 | 3 | 2 | 10 | 1 | 0 | 0 | 0 | 1 |
| $p\bar{q}$ | 1 | 0 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 3 |
| $\bar{p}q$ | 1 | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 1 | 2 |
| $\bar{p}\bar{q}$ | 4 | 3 | 2 | 5 | 14 | 12 | 10 | 11 | 10 | 43 |

Consequent ($F = 16.28$, $p < .001$) and Rule-forms, Antecedent and Consequent ($F = 12.84$, $p < .01$; all ratios were tested on conservative degrees of freedom). For the interpretation of these effects, one may refer to Table 25 and inspect the means for each rule and rule-form. The data are pooled across groups because there was no hint of any effect of materials. Looking at Table 25 we can see that the order of difficulty observed in the last experiment of AA < AN = NA < NN is upheld on the IT and OI forms, but that the pattern is different on the EO form, where it is AA < NN < AN = NA.

Verification times were subjected to a 2 x 3 x 4 x 4 (Groups x Rule-forms x Rules x Matching cases) analysis of variance (see Appendix 6). There were significant main effects of Rule-forms ($F = 5.54$, $p < .05$), Rules ($F = 20.10$, $p < .001$) and Matching case ($F = 5.29$, $p < .05$), and a significant interaction between these three factors ($F = 4.71$, $p < .05$; all assessed on conservative degrees of freedom). The three lower-order interactions between these factors were also significant. The most meaningful course in interpreting these effects is to examine the mean latencies for each rule-form separately; these are set out in Table 26, and the orderings about to be mentioned, which should at this stage be considered as approximations, were arrived at with the aid of Scheffe comparisons within each rule-form. On the IT form the order of difficulty of the rules is not well distinguished, but seems to follow the order AA = AN = NA < NN, with the order of the matching cases being $pq = \overline{pq} < p\overline{q} = \overline{p}q$. These orders are similar to those observed in Experiment 7 on comparable rules, and once again the shortest verification times are to the double-mismatching ($\overline{pq}$) case. The situation is different with the OI form: there is little variation among the rules, although NN again incurs the longest times. The order for matching cases is $pq = p\overline{q}$

TABLE 25 Mean Comprehension times pooled across groups for the four rules in each rule-form. Data from Experiment 8.

| | Rules | | | | |
|---|---|---|---|---|---|
| Rule-forms | AA | AN | NA | NN | Mean |
| If then | 4.80 | 6.51 | 6.73 | 8.31 | 6.59 |
| Only if | 6.12 | 7.83 | 8.55 | 11.81 | 8.58 |
| Either or | 5.20 | 9.48 | 10.22 | 7.67 | 8.14 |
| Mean | 5.37 | 7.94 | 8.50 | 9.26 | 7.77 |

TABLE 26 Mean Verification times for the three rule-forms pooled across groups and ordered in terms of rules and matching values. Data from Experiment 8. Times in seconds.

'IF THEN'

| | Matching value | | | | |
|---|---|---|---|---|---|
| Rules | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ | Mean |
| AA | 2.98 | 3.20 | 6.72 | 5.43 | 4.58 |
| AN | 3.99 | 4.75 | 6.88 | 4.62 | 5.06 |
| NA | 5.08 | 8.28 | 4.65 | 4.89 | 5.73 |
| NN | 8.70 | 12.17 | 8.98 | 4.52 | 8.59 |
| Mean | 5.19 | 7.10 | 6.81 | 4.86 | 5.99 |

'ONLY IF'

| | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ | Mean |
|---|---|---|---|---|---|
| AA | 3.26 | 5.48 | 11.10 | 7.35 | 6.80 |
| AN | 5.10 | 4.16 | 8.89 | 12.57 | 7.68 |
| NA | 4.90 | 8.16 | 5.30 | 9.16 | 6.88 |
| NN | 6.84 | 8.31 | 13.87 | 5.86 | 8.72 |
| Mean | 5.02 | 6.53 | 9.79 | 8.74 | 7.52 |

'EITHER OR'

| | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ | Mean |
|---|---|---|---|---|---|
| AA | 8.37 | 4.91 | 5.24 | 4.43 | 5.74 |
| AN | 8.31 | 7.86 | 7.15 | 4.67 | 7.00 |
| NA | 10.27 | 10.65 | 6.72 | 5.97 | 8.40 |
| NN | 9.54 | 6.30 | 5.46 | 8.35 | 7.41 |
| Mean | 9.12 | 7.43 | 6.15 | 5.86 | 7.14 |

< $\overline{pq}$ = $\overline{\overline{pq}}$ - the double mismatching instance is by no means the fastest here. In the case of the EO form the AA rule is the easiest, with little difference between the other three, and of the matching cases, pq takes much the longest to verify, with the others about the same, and $\overline{\overline{pq}}$ again the shortest.

In the VT analysis there was some effect of materials: the group factor interacted with both rules ($F = 4.09$, $p < .01$) and matching values ($F = 5.11$, $p < .01$). The relevant means are displayed in Table 27. As in Experiment 7, there is no overall materials effect in the VTs, but rather pattern differences in the rule and matching factors between the groups. These effects must run across rule-forms, as materials do not interact with this factor. Thus in the Abstract group the NN verification latencies are longer than the others, whereas there is no such trend in the Thematic group; in the Abstract group there seem to be no appreciable differences between the matching cases, but in the Thematic group the latencies on the $\overline{\overline{pq}}$ case, and to a lesser extent the pq case, are substantially the shorter.

## DISCUSSION

Interpreting these results, especially the latency data, could easily descend into a vision of too many trees and not enough wood. The latencies will therefore be discussed primarily in terms of the overall effects observed in them, and comments on cell means kept to a minimum. In doing this one runs the risk of appearing vague, but it must be remembered that the latencies were essentially a supplementary measure in this and the previous experiment, and that the practical usefulness of high-order interactions arising from multifactorial analyses of variance from an N of 32 is debatable. The

TABLE 27 (a) Mean Verification times for the two groups ordered in terms of matching value and pooled across the other factors. Times in seconds. Data from Experiment 8

| | | Matching value | | | | |
|---|---|---|---|---|---|---|
| Groups | | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ | Mean |
| Abstract | | 6.58 | 6.84 | 7.06 | 7.10 | 6.90 |
| Thematic | | 6.31 | 7.19 | 8.10 | 5.87 | 6.87 |
| | Mean | 6.45 | 7.02 | 7.58 | 6.49 | 6.89 |

(b) Mean VTs for both groups ordered in terms of rules and pooled across the other factors. Data from Experiment 8.

| | | Rules | | | | |
|---|---|---|---|---|---|---|
| Groups | | AA | AN | NA | NN | Mean |
| Abstract | | 5.39 | 6.09 | 6.82 | 9.26 | 6.89 |
| Thematic | | 6.02 | 7.06 | 7.18 | 7.23 | 6.88 |
| | Mean | 5.70 | 6.58 | 7.00 | 8.25 | 6.89 |

major emphasis of interpretation is on the response frequencies, not only because in reasoning research it is the solutions which people arrive at rather than the time they take to arrive which is of primary interest, but also because some revealing trends in the frequency data have emerged. The implications of the current findings generalise beyond the present experimental setting, and these general implications will be discussed in due course, but for the moment we need to consider the results as they relate to previous truth-table research.

The 'if then' rules.

The results of the matching analyses both within and between the groups, and an inspection of Tables 23 and 24 along with Table 19, show that the results for the IT rules in Experiment 8 provide an almost exact replication of the findings of Experiment 7. This is continued to a lesser extent in the latencies, where the overall mean CTs and VTs are very close. However, the materials effect on CTs in Experiment 7 has not been repeated: there was no evidence in Experiment 8 that any of the thematic rules were easier to comprehend than the abstract rules. This may be due to the slight change in wording, but this is unlikely as wording changes have been found to have little effect on the Selection task, and there was no effect on actual responses. It is more likely that the embedding of the IT rules among eight other sentences was responsible, subjects having got into a more uniform rhythm of responding under which only gross differences would emerge. Although there was no significant interaction of materials with other factors in the VT analysis, it is interesting to note that the mean VTs for the two groups take similar patterns in Experiment 8 as in Experiment 7 (see Tables 27a and 21a). There is little difference between the matching items in

the Abstract group, but a marked speeding of evaluation of the pq and $\overline{pq}$ instances in the Thematic group.

The 'only if' rules.

The results for this alternative form of the conditional show a general similarity to those for the IT form, but there are some differences which should also be noted. In the response frequencies the results of the matching analyses are much the same, with the greater rejection of the $\overline{pq}$ case as irrelevant under thematic materials. In the latency analysis however there are some substantial differences. Firstly, the comprehension times are on average a full two seconds longer for the OI rules than for the IT form, with a particular difference on the doubly-negated rule; indeed, some of the subjects remarked during debriefing how difficult it was to make sense of a 'not p only if not q' sentence. Similarly, the VTs form a distinctly different pattern, with generally longer times and no suggestion of the rapidity of responding to the $\overline{pq}$ instance observed on the IT rules - a fact which must constrain any general conclusions about this. Taken together, these results indicate that the 'only if' form is probably a less natural expression of a conditional relationship than 'if then', and that negation can cause especial difficulties with this form which are not experienced with IT sentences. Similar conclusions arose from the study of these two expressions by Evans & Newstead (1977), who also showed that the OI form took a more natural part in expressing a reverse temporal order of antecedent and consequent. There was no such specification of temporality in the present experiment, so perhaps the OI form acquires some of its extra difficulty when used outside this particular context.

The 'either or' rules.

There is not a great deal of previous research with which the present results can be compared; the nearest relative to the present experiment is an in press study by Evans & Newstead. Although they used construction and evaluation procedures, the latter without an 'irrelevant' response category, they did use the latency measures, and also left the judgement of inclusive v. exclusive disjunction to the subjects. We therefore consider three important facets of the EO data in turn: frequencies, latencies, and inclusive/exclusive classification.

Firstly, then, the response frequencies. It has been contended before that the disjunctive rule-form is immune to matching bias (e.g. van Duyne, 1973), and this is confirmed in Experiment 8. Evans & Newstead (in press) came to the same conclusion, and remark, along with Johnson-Laird & Tagart (1969), that this was probably due to a lack of use of the 'irrelevant' category in considering the disjunctive. Certainly, there were fewer cases in Experiment 8 judged irrelevant to the EO rules (11%) than to the IT (24%) or OI (18%) rules here in the Abstract group; the treatment of the thematic rules seems to have been radically different. The relative consistency of responding observed by Johnson-Laird & Tagart on an abstract NA disjunctive is entirely absent here: no one classification pattern to this rule appeared more than three times. Of course, the most striking feature of the frequency data is the difference in the patterns of 'irrelevant' responses between the two types of materials, with the $\overline{pq}$ case being ruled out in the same way as it was on the conditional rules. The difference is all the more marked because the pattern of 'irrelevants' in qualitatively changed

in the two groups - it is not that there is simply an apparent change in responding to one particular contingency, as seemed to be happening on the conditionals. With thematic disjunctives, it seems that the double-mismatching instance is the only one considered irrelevant, and it is considered irrelevant most of the time. The similarity of treatment of this instance on all three rules, and the fact that it quite plainly does not reflect a simple matching bias effect on the disjunctives, has general implications for a theoretical account of the treatment of the thematic task. This, and the difference between the materials, will be enlarged upon after examination of the other aspects of the disjunctive data, latencies and the inclusive/exclusive classification, to which we now turn.

It has often been noted that singly-negated disjunctives give rise to fewer logically correct solutions than do doubly affirmative or negative disjunctives (e.g. Roberge, 1976; Johnson-Laird & Tridgell, 1972). This could arise from an operational difficulty due to, for instance, denying a negative with an affirmative, or to the singly-negated rules being simply more difficult to understand. Evans (1972c) argues for the latter, and a ready prediction from this argument is that comprehension times to the AN and NA disjunctives should be longer than those for the AA and NN rules. The Evans & Newstead study cited above looked at this, and indeed found an interaction between negation of the first and second components, but only to the effect that latencies to the AA rules were shorter than to the other three. This is not very surprising, and can be taken as only partial confirmation of the initial hypothesis. However, the CTs obtained in the present study provide stronger

confirmation: from Table 25 we can see that the AN and NA rules really did take longer to comprehend than the AA and NN, under both types of materials, and that AA times were shorter than NN times. Solution latencies (VTs) form the same pattern in this experiment as in Evans & Newstead's: the AA sentence records the shortest time, with little difference between the others. In terms both of understanding and evaluation then, negation seems to cause particular difficulties with the disjunctive, and a single negative can make these difficulties acute; thematic materials do not alleviate the problems.

Whether the disjunctive is taken to carry an inclusive (p or q or both) or exclusive (p or q but not both) connotation seems to depend on whether a given author is writing from a logical or linguistic standpoint. The logicist point of view is that the disjunctive should be considered inclusive unless specified otherwise (e.g. Wason & Johnson-Laird, 1969) and perhaps not even then (Ennis, 1976), while the linguistically oriented psychologist takes exactly the opposite view, as expressed by Fillenbaum (1974a): "in natural language it may...be quite difficult to interpret 'or' in an inclusive sense". Which of these views will be the correct one when unqualified disjunctives are evaluated is an empirical question, -which can be answered by examining both the present data and those of Evans & Newstead (in press). The answer centres on the TT logical case: it verifies an inclusive, since this allows the occurrence of both items together, but falsifies an exclusive, since this prohibits their co-occurrence. Evans & Newstead, using abstract materials, found a majority of subjects classifying the TT case as true, ie. adopting the inclusive classification, but in the present study

there was a clear preference for the exclusive; nine out of sixteen subjects in the Abstract group classified the TT case (on the AA rule) as false and only three classified it as true, and the preference was even clearer in the Thematic group, where 13 of the 16 subjects used the exclusive classification. The conclusion, a somewhat weak one perhaps, must therefore be that both inclusive and exclusive classifications are open to the subject, the direction of the choice being clarified by context. It is worth noting that the verification latency to the TT case on the AA disjunctive in the present study was appreciably longer than those to the other cases, so perhaps subjects spent some time agonising over whether the sentence allowed or prohibited the co-occurrence of its constituents.

## CONCLUSION

We can now summarise the main findings of Experiment 8. Firstly, the results from the IT rules provide a good replication of the results of Experiment 7, except for an interactive effect in the comprehension times that did not reappear. The different expressions of the conditional made little difference to the reasoning responses to them, a result which has been found before (see Wason & Johnson-Laird, 1972). Taking the two conditionals in Experiment 8 (IT and OI) together, we can see that the response frequencies to the abstract rules accord closely with the patterns previously observed by Evans (1972b, 1975; Evans & Newstead, 1977) when negatives are used in the rules: there is an increasing tendency for instances to be classified as irrelevant to the extent that they do not match the items named in the rules. In the thematic rules there are similar basic patterns, except for a much increased ruling out of the

double-mismatching ($\overline{pq}$)case as irrelevant. There is no indication that thematic materials lead to more logical evaluations, or even that they make the conditionals easier to understand (CT) or reason with (VT). In the case of the EO rules, it was found that the preferred classification of the TT logical case was in accord with an exclusive interpretation of the disjunctive, and that this preference was most clearly marked under thematic materials. The predicted difficulty due to negation was observed under both types of content (VT), with the singly-negated rules taking substantially longer to understand than the AA rule, with NN in between (CT). It is difficult to come to a meaningful assessment of the proportions of logically correct responses across rules since the classification patterns differ markedly between the two groups, an effect seen most graphically in the responses to the $\overline{pq}$ case. Here the pattern observed in the Abstract group, and to a lesser extent in the Evans & Newstead study, is replaced entirely in the Thematic group. These differences, and the similarity of the $\overline{pq}$ response patterns to those seen on thematic conditionals, provide some cluses as to the nature and source of the varying treatment of the two types of content. This theoretical account will be undertaken in the next chapter.

## PART THREE: GENERAL DISCUSSION

### CHAPTER 8



Tables

A theoretical account of truth-table task performance

In the next few pages a theory of truth-table task performance will be developed, based primarily on the data from the Thematic groups in Experiments 7 and 8. In much of this discussion, although it is the data from the latter experiment which will be referred to explicitly, the results of Experiment 7 are implicit in it, as they were repeated almost exactly and are covered by any points made about Experiment 8. The theory is based on the Thematic data because it is here that new and interesting trends have emerged which can be used as indicators to distinguish two psychological formulations of truth-table performance, which cannot be distinguished on the basis of the results from the Abstract group. The first of these formulations is Evans' (1972c, 1977b) well-known conception of the competition in reasoning performance of two cognitive (or statistical) factors: interpretative and operational tendencies. These correspond respectively to the logical requirements of the task and to non-logical response factors such as matching. The second formulation, and the one that will be urged here, is that the distinction of interpretation and operation may, at least in the truth-table task, be artificial, that 'matching bias' may be a misnomer although the behaviour it refers to is genuine enough, and that truth-table performance can be viewed as an active attempt on the subjects' part to construct treatments of the materials with which they have to work. This view will also be applied to other reasoning situations.

Let us begin by looking again at the results of the last two experiments. In brief, it was found that Abstract subjects' truth-table classifications accorded with previously found patterns of responding with, on conditionals, a tendency to rule out as irrelevant those instances which mismatched the items named in the rules (matching

bias); and that this tendency was greater when thematic materials were used. In the case of disjunctive sentences, this increase represented a total change between the materials - abstract disjunctives were immune from the matching effect. This is an oversimplified account of what actually happened, as we shall shortly see.

There are two ways of characterising matching behaviour psychologically. Firstly, there is Evans' (1972b, c) original formulation: that the classification of mismatching items as irrelevant is due to a pure response tendency which cuts across an otherwise chiefly logical appraisal of the instances. This is demonstrated by the application of negation to conditional rules; in the ordinary affirmative (AA) sentence the ruling out of mismatching cases coincides with Wason's idea of a 'defective' truth-table, in which 'irrelevant' is a third value besides 'true' and 'false'. However, another way of looking at the behaviour observed with negated rules is to view it as arising from re-interpretations of the rules rather than responses to the instances: if subjects were ignoring the negatives in the rules, the matching cases would have the same logical values across all rules, and the 'irrelevant' responses would represent the application of the same defective truth-tables to these recast rules. Evans assumes that negatives <u>reflect</u> a response bias, i.e. that there is no interaction between negation and interpretation/operation, but the alternative view is that negation <u>creates</u> different treatments, and that this is what is manifested in the 'matching' data. We should be able to distinguish between these alternative explanations by recourse to procedures which affect the behaviour, and that is just what the thematic materials in Experiment 7 and 8 did, and do.

The original formulation of matching bias runs into trouble on two counts when the Thematic results from Experiments 7 and 8 are examined in more detail. Firstly, it should be noted that there are three mismatching cases for a two-component conditional or disjunctive sentence: the TF ($p\bar{q}$), FT ($\bar{p}q$) and FF ($\bar{p}\bar{q}$) instances. Any theory which attempts to account for the effects observed under thematic materials in terms of an increased matching response tendency would have to predict some increase in 'irrelevant' responses to all three cases. However, in the last chapter we saw that the increases in the antecedent and consequent matching effects on the conditionals were due to increased 'irrelevant' responses to the $\bar{p}\bar{q}$ case alone - 'irrelevant' responses to the $p\bar{q}$ and $\bar{p}q$ cases were no different between the groups. Secondly, there are the disjunctive rules. Abstract disjunctives have been found to be immune from the matching effect before, and were found to be immune again in Experiment 8. However, the thematic disjunctives produced highly significant matching effects, again due entirely to an overwhelming rejection of the $\bar{p}\bar{q}$ case as irrelevant relative to the other cases. Matching bias must account both for its one-sidedness in this situation, and for its creation out of nothing. There is a simpler and more plausible alternative to the stretching of matching bias, and it may be illustrated by looking at the modal classifications which subjects gave to each instance in Experiment 8.

These classifications are given in Tables 28 and 29. In Table 28 they are arranged in terms of logical contingencies pooled across matching values, to illustrate the effects of logic and matching; in Table 29 the ordering is the other way round, with matching values pooled across logical cases. Both Abstract and Thematic data are summarised in this way, but for the moment our concern is with the

Thematic data only.

If we first take the 'logical' Table 28, two things are apparent straight away: the dismissal of the $\overline{pq}$ case running diagonally from bottom left to top right across rules and instances, and what appear to be the underlying patterns on which this response is superimposed. For the conditionals, these underlying modal evaluations are always for TT as true, TF as false, FT as false, and FF as true. On the disjunctives, response patterns are split between the singly negated rules (AN and NA) and the other two (AA and NN), with the former similar to the evaluations for the conditionals (i.e. equivalence with $\overline{pq}$ irrelevant) and the latter with a reading as exclusive disjunction, again with $\overline{pq}$ irrelevant. Taken together, this behaviour is somewhat paradoxical: the subjects seem quite clear that the thematic conditionals in this experiment were rules of equivalence, but equally clear that a contingency neither of whose components appeared in the rule could have nothing to do with that rule. This seems particularly strange in the case of the NN rule, 'If not p then not q' or 'Not p only if not q'. Can subjects really believe that 'not-p and not-q' - the TT case - is irrelevant to this rule, as they did in 75% of cases here? Furthermore, they seem fully aware of the role of the TT case both in the abstract version of this task, where it was recognised as verifying the rule 78% of the time, and in Evans' (1972b) construction truth-table task, where 92% of subjects immediately gave the TT case when asked to compose an instance to verify an NN rule. The confusion is no less on the disjunctive rules, where the subjects seem to use at least two truth-tables but are just as certain of the irrelevance of the $\overline{pq}$ case, which is the only one seen most often as irrelevant, and is seen so 67% of the time.

This confusion is alleviated if we consider the second explanation of performance on the truth-table task, which is illustrated

TABLE 28 Modal classifications of each logical contingency in Experiment 8. T = true, F = false ? = irrelevant

| | Abstract | | | | Thematic | | | |
|---|---|---|---|---|---|---|---|---|
| **IF THEN** | | | | | | | | |
| | Logical contingencies | | | | | | | |
| Rules | TT | TF | FT | FF | TT | TF | FT | FF |
| AA | T | F | F | ? | T | F | F | ? |
| AN | T | F | ? | T/? | T | F | ? | T |
| NA | T | F | F | T | T | ? | F | T |
| NN | T | F | F | T | ? | F | F | T |
| **ONLY IF** | | | | | | | | |
| | TT | TF | FT | FF | TT | TF | FT | FF |
| AA | T | F | F | ? | T | F | F/? | ? |
| AN | T | F | ? | T | T | F | ? | T |
| NA | T | ? | F | T | T | ? | F | T |
| NN | T | F | F | T | ? | T | F | T |
| **EITHER OR** | | | | | | | | |
| | TT | TF | FT | FF | TT | TF | FT | FF |
| AA | F | T | T | F | F | T | T | ? |
| AN | T | T/F | T | F | T | F | ? | T |
| NA | T | T | F | F | T | ? | F | T |
| NN | F | T | T | F | ? | T | T | F |

TABLE 29 Modal classification of each matching contingency in Experiment 8. T = true, F = false, ? = irrelevant.

| | Abstract | | | | Thematic | | | |
|---|---|---|---|---|---|---|---|---|
| **IF THEN** | | | | | | | | |
| | Matching contingencies | | | | | | | |
| Rules | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ |
| AA | T | F | F | ? | T | F | F | ? |
| AN | F | T | T/? | ? | F | T | T | ? |
| NA | F | T | T | F | F | T | T | ? |
| NN | T | F | F | T | T | F | F | ? |
| **ONLY IF** | | | | | | | | |
| Rules | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ |
| AA | T | F | F | ? | T | F | F/? | ? |
| AN | F | T | T | ? | F | T | T | ? |
| NA | F | T | T | ? | F | T | T | ? |
| NN | T | F | F | T | T | F | T | ? |
| **EITHER OR** | | | | | | | | |
| Rules | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ | pq | $p\bar{q}$ | $\bar{p}q$ | $\bar{p}\bar{q}$ |
| AA | F | T | T | F | F | T | T | ? |
| AN | T/F | T | F | T | F | T | T | ? |
| NA | F | F | T | T | F | T | T | ? |
| NN | F | T | T | F | F | T | T | ? |

by Table 29. Here the modal classifications are ordered according to their matching rather than logical status. In this table there are generally only two 'truth-tables' throughout: defective truth-tables for equivalence and exclusive disjunction (on the OI rule-form there is some evidence for a treatment of the AA and NN rules as implications). This table expresses the essence of the alternative explanation of truth-table performance: that the 'irrelevant'responding to mismatching cases reflects a tendency to apply similar truth-tables to reformulations of the rules. The ordering in Table 29 assumes that subjects are responding to rules of equivalence and exclusive disjunction and ignoring negatives, such that the matching cases are the logical cases for those rules. Table 30 lists the reformulated rules which the subjects seem to be applying. The AA and NN conditionals (noting the variations in the OI form) therefore follow the pattern of defective equivalence, while the AN and NA conditionals and all the disjunctives are treated as unnegated defective exclusives. All truth-tables are defective - the $\overline{pq}$/FF case is regarded as irrelevant to all rules.

Surely this apparent circumventing of the negative is quite irrational? In this case, the subjects can be excused their treatments of the rules - they were in fact acting quite sensibly here. The question of whether a conditional can justifiably be treated as an implication or an equivalence is, as we have seen, controversial (see Chapter 3). It can be specified, of course, but when no such guide is given it is almost an open question which treatment to adopt; certain contents which seem to favour an equivalence interpretation can be listed (e.g. Wason & Johnson-Laird, 1972), but this is not to specify the sufficient conditions. It is quite plain that the food-and-drink

TABLE 30 Recast rules used in the Thematic group in Experiments 7 and 8.

CONDITIONALS

| | Original | Recast |
|---|---|---|
| AA | If p then q<br>p only if q | If and only if p then q |
| AN | If p then not q<br>p only if not p | Either p or q, but not both |
| NA | If not p then q<br>Not p only if q | Either p or q, but not both |
| NN | If not p then not q<br>Not p only if not q | If and only if p then q |

DISJUNCTIVES

| | Original | Recast |
|---|---|---|
| AA | Either p or q | Either p or q, but not both |
| AN | Either p or not q | Either p or q, but not both |
| NA | Either not p or q | Either p or q, but not both |
| NN | Either not p or not q | Either p or q, but not both |

rules here were regarded by most subjects as (defective) equivalences, as the classifications on the AA rules show. The NN rules were treated in the same way. This implies that the subjects were ignoring the negatives in processing the NN rule, and indeed they might have been. Fortunately for them, the logical truth-tables for AA and NN equivalences are identical and symmetrical - no great logical error ensues from ignoring the negatives in an NN equivalence, even when applying the defective truth-table. A similar process seems to have governed the evaluation of the AN and NA conditional. Here the classification patterns are identical with that for unnegated exclusive disjunction (with $\overline{pq}$ irrelevant). Again this can be justified: singly-negated equivalences and unnegated exclusives are logical isomorphs, a fact noted by Wason & Johnson-Laird (1972) in their discussion of 'disguised disjunctives' (pp. 61-2). They were referring to NA conditionals, but in the present experiment it seems that the subjects also regarded 'If pork then not wine' (AN) to mean the same thing as 'pork or wine'. This analysis resolves the question of why only the $\overline{pq}$ case was rejected: with the application of similar truth-tables across recast rules we would only expect the one case to be treated in this way.

The subjects seem, then, to have hit upon a way of treating all the conditionals as their nearest unnegated logical correspondents - they have seemingly striven to avoid negation, but remained quite rational in so doing. One cannot say the same for their treatment of the disjunctives, where the defective exclusive pattern persists throughout. This shows that the strategy adopted was one of a wholesale ignoring of negatives. Thus the difficulty which has repeatedly been found in reasoning with negated disjunctives (e.g. Evans, 1972c; Roberge, 1976, 1977; Evans & Newstead, in press) seems to have effectively defeated the subjects here. They did not respond by guess-

work - their response distributions were not random - but by making the best sense of these incomprehensible sentences as they could: treating them as if the negatives were of no account. The problem of explaining the appearance of matching bias in the thematic rules where it did not exist in the abstracts is removed in favour of an account stressing the application, albeit logically erroneous, of a truth-table.

The abstract task

Before enlarging on this interpretational account of performance in the present thematic task, we need to go back and consider the differences between this performance and that observed on the abstract form of the task. In fact, between the abstract and thematic conditionals the differences are not great, as one can see by inspecting Tables 28 and 29. The Abstract response profiles, reflected in the modal classifications of Table 29, are, so to speak, less defective - $\overline{pq}$ is not ruled out as irrelevant so often, but otherwise the patterns are much the same as for the Thematic group, and where the $\overline{pq}$ case is not mostly considered irrelevant, the modal responses are in the 'right' directions, i.e. in line with the formal correspondent of the particular defective truth-table. At this point it is worth remembering that the above account of the re-interpretation of given rules is one of tendencies to re-interpret. Just as Evans is sometimes wrongly accused of claiming that all reasoning is matching, so it would be wrong to infer from this account that the argument is that all reasoning proceeds by re-interpretation. One would hardly expect a re-interpretation where the one assumed by a logician or a reasoning researcher was sufficient for a subject to proceed with. Indeed, one can see some evidence of a competing tendency to adhere,

for instance, to the 'real' conditional structure of the AN and NA rules: on both rules, there are more evaluations of an instance as true where this instance corresponds to the logical TT case than when it corresponds to the FF case. Thus the thematic materials, relative to the abstracts, seem to have clarified for the subjects the treatments they should give to the sentences; there is less variability both between and within rules. Where such clarification is lacking, as in the case of the abstract rules, it appears that some subjects will follow the conditional structure and some the recast structure, but there is not the concensus one way or the other which the thematic materials evoke. There is, therefore, a thematic materials effect in the truth-table task, but towards particular interpretations rather than greater adherence to formal logic or to matching. One could, perhaps, take Evans' (1977b) statistical model of Selection task performance and adapt its parameters to refer to this behaviour: one could say that there are probabilities of different classification patterns (truth-tables) emerging in different contexts. For instance, the abstract task used here is less linguistically based than the thematic task, and a straightforwardly concrete task might yield still different reasoning patterns (see e.g. Legrenzi, 1970; Rips & Marcus, 1977). The test of this idea is to vary both rule content and problem context; beyond this, one cannot say much more about the materials effect seen here.

The difference between abstract and thematic responding and between confusion and clarity is seen most strongly in the data from the disjunctive rules. Here the abstract responses do not vary around the logic of the problem at all in the case of the AN and NA rules, and no kind of truth-table seems to have been used in them. Evidently, these rules make no sense at all to the subjects, a conclusion at which previous researchers arrived some time ago. Thematic materials make

a dramatic difference: the subjects may not follow the formal structure, but there is a large measure of agreement among them as to the re-interpretations which are to be followed.

Equivalence, exclusion, and defective tables

The question was touched upon earlier of why the subjects should choose equivalence and exclusive treatments of the sentences in these experiments, rather than implications and inclusives. Some other questions are related to this and should also be considered: why should the truth-tables in the Thematic group have been defective, and why did the subjects apparently seek to minimise the role of negation by recasting the rules in unnegated forms? Past research provides some suggestions.

As far as the interpretation of the conditional and disjunctive goes, we have seen in preceding chapters how opinion differs on the legitimacy or otherwise of the various possible readings of these sentences. It was noted that logically oriented psychologists tend to assume the inclusive interpretation of the disjunctive, and those with a linguistic bent, the exclusive; the same seems to follow for implication and equivalence treatments of the conditional. Although the lack of correspondence between the logical and linguistic expressions of the conditional has been noted by logicians (e.g. Strawson, 1952), it has usually been assumed that subjects should interpret 'If p then q' as implication, and some surprise is not unknown when they do not (e.g. Taplin & Staudenmayer, 1973; see Chapter 3). Geis & Zwicky (1971), writing from a purely linguistic standpoint, argue against this, and place the weight of their emphasis on the other end of the scale: conditionals, they contend, are normally taken to express equivalence unless the context specifies otherwise. Truth-table experiments rarely make any such specification, and Geis & Zwicky's dictum seems to have been borne out empircally, because in

the history of truth-table research since Johnson-Laird & Tagart (1969), the FT logical case, the critical contingency in differentiating implication and equivalence evaluations, has only seldom been classified as 'true' (i.e. implication) by subjects (see also Evans, 1972b, 1975; Evans & Newstead, 1977; Rips & Marcus, 1977). There is some evidence for defective implication, with the FT case classified as irrelevant, but the most common category for this instance is 'false', which is in line with an equivalence treatment. The present study continues this line.

In the case of the disjunctive, the most 'natural' treatment is more equivocal, with as much evidence for the inclusive as for the exclusive. The same points about context would apply to this, and in the present experiment the materials used seem to have been taken as clearly implying an exclusive reading, in the same way that they seem to dictate equivalence in the conditionals. The two are probably close psychological as well as logical relatives: we saw how there is evidence in the response patterns to the singly-negated conditionals that the two are indeed closely related cognitions.

The role of the $\overline{pq}$ case and the reasons for the subjects' attempt to neutralise negation are best explained by taking a step back from the experimental setting and considering how such sentences would ordinarily be generated and applied in natural language. The (defective) equivalence conditional requires the establishment of both the antecedent and consequent; the establishment of neither of them renders the statement vacuous. Why the $\overline{pq}$ case should be irrelevant to an exclusive disjunction is less obvious: as the statement is that there should be one thing or the other, the absence of both seems to constitute a plain refutation. However, in ordinary language would one expect a statement of strict alternation to be made at all if

there were any likelihood of the absence of both alternatives? It would seem that exclusive disjunction is an expression to be used when it is given that there will be one alternative present - the question is which one - and once again the statement becomes meaningless in the presence of neither.

We saw in Chapter 3 how Rips & Marcus' (1977) Suppositional theory of conditional interpretation claimed to account for the defective truth-table, and what they say has some relation to the arguments above. It will be recalled that their idea was that the truth value of a conditional statement rested on consideration of a supposition from which the statement was derived, and that this supposition consisted of an addition of the current data base - things considered relevant to the statement - and the 'seed proposition' embodied in the antecedent. This hypothesis works best with implicative statements; to extend it to equivalences would require the assumption of two 'seed propositions', for the antecedent and consequent, since the rule is bi-directional. Inasmuch as both the present account and Rips & Marcus' theory address the questions of context and the defective table, the two approaches are related. Rips & Marcus also provide some clues as to why the content of the present thematic task should have lent itself so readily to equivalence (and exclusion) response patterns. They found that a correlational relationship between the two terms in a conditional led to a preponderance of equivalence responses; perhaps the foods and drinks in the rules used here were seen as correlated, as foods and drinks tend to be in real life. It is recognised that both these hypotheses veer towards circularity, and also that any theory proposing context effects and content relationships after the fact must run the same risk. This question will be confronted again a little later.

The subjects' treatment of negated rules, in the case of the conditionals as their nearest unnegated equivalent and in the case of the disjunctives as if the negation did not exist, bears some relation to findings in other fields. There are many recorded instances in psycholinguistic research where subjects, given the chance, will recall, reconstruct, or evaluate given negated sentences in simpler, unnegated forms. Mehler's (1963) finding that recall was more accurate for active-affirmative sentences than for negatives or passives has been extended to incidental learning (Cornish & Wason, 1970) and the recall of instructions in a natural situation (File & Jew, 1973). Fillenbaum (1974b) has shown that this apparent primacy of the active-affirmative may extend into an active process of reduction in comprehension. He found that subjects tended to reduce sentences in a paraphrase task to forms which were not only syntactically simpler but semantically simpler as well, as if they were 'correcting' what the experimenter was trying to say but expressing badly. Interestingly, one such paraphrase was from a negated to an unnegated disjunction, changing the logical meaning of the sentence in the same way that the subjects did in Experiment 8. Fillenbaum aptly calls this "pragmatic normalisation", and, to be generous, this may be what the subjects were attempting in ignoring the negatives in the EO rules here. A negated disjunction is such a bad expression that it has to be reduced to a 'normalised' form, in this case the unnegated rule. There really seems to be no such thing as a negated disjunctive, at least when making rules about foods and drinks or letters and numbers. The process of normalisation also provides an attractive account of what the subjects did with the conditional

rules: why use a negated conditional when an unnegated disjunctive or equivalence amounts, pragmatically, to the same thing?

The view of the processing of rules and truth-table cases propounded so far is a wholist one, in which the separation of interpretation and operation is seen as artificial. Evans (1972c, 1977b) hints at this when writing of the interaction between response biases and logical considerations, but does not go so far as the present discussion, which argues that not only will the meaning of the sentence influence the evaluation of the instance, the presence of an instance may influence the meaning of the sentence. The two are most obviously inseparable in the TT (as true) and TF (as false) cases, which are almost shorthand expressions of the meaning of 'If p then q'. All meanings of this sentence must include these cases; other meanings are reflected in the values given to the FT and FF cases. These cases, though, may in themselves affect the subject's assessment of the sentence - he might not even confront the question of whether his conditional is or is not to imply its converse until he confronts these cases. If FT looks like a plausible falsifier, for whatever reason, he might be deflected towards a judgement that it does falsify. The question asked of interpretation and operation - which came first - is the same one asked of the chicken and the egg, and the answer to both is, of course, neither: they evolve together. In comprehending a logical expression such as 'If then' or 'Either or', one is engaging in a truth-table task, and the other factor influencing this process is context, the situation in which the process occurs. To take a trivial example, and pay one last visit to our tigers, we know that

'If it is a tiger then it has stripes' is an implication, because FT does not falsify it. However, uttered in the context of a discussion about large carnivores of the Indian sub-continent, it is surely an equivalence - we know that tigers are the only large Indian carnivores with stripes, so a non-tiger with stripes (FT) is now a falsifying instance.

In short, then, the cognitive activity in a truth-table task consists of a parallel processing of sentence and instance. The distinction between interpretation and operation implies a serial process (cf. Evans & Newstead, in press, who make a similar point), and the splitting of latencies into comprehension and verification times must assume serial processing to have any validity. The argument must therefore be (assuming equivalence!) that if the processing of sentence and instance is parallel, then the CT-VT procedure is not valid; it is the main reason why latencies have not been afforded too much attention in this discussion. The splitting procedure, on this argument, might also have affected the actual responses, though such effects would probably not have been serious: not only could the subjects undertake several re-evaluations of each rule, they could also establish a general response strategy over the sixteen rules.

Applications: construction and Selection tasks

The first test of any model is to assess its generality by comparing its explanatory merits in alternative versions of the situation from which it first arose, and in different situations in which there is reason to believe that similar things might be happening, especially if similar linguistic materials are used. As regards the first question, some data from a different form of

the truth-table task is already available, in Evans' (1972b) report of an experiment in which subjects had to construct their own instances rather than evaluate ones they had been given. One should clearly expect similar trends in the data as those observed in the present experiments, and indeed there are. However, there are two snags in this apparently cosy situation. Firstly, Evans' task was an abstract task; one would expect more variability of responding on an abstract task than on equivalent thematic task, but there are no such thematic results on which to base such a comparison. Secondly, and on a more important point, the predictions of matching and re-evaluation cannot be distinguished in this one experiment. One could predict, say, that there will be fewer $\overline{pq}$ instances than others constructed overall, since this forms the 'defective' part of the truth-tables described above; or that there will be more of a particular case constructed in a particular way when truth-table cases under both interpretations (avoding and accounting for negatives) coincide. Matching bias predicts the same statistical effects through the competition of logic and matching: the $\overline{pq}$ case will not be seen as relevant because it mismatches, and the 'strong' (i.e. non-defective) truth-table cases are also those on which the least effect of matching bias would be expected, pq and $p\overline{q}$. One could split the matching and re-interpretation theories by introducing something which shifts responding and seems to favour one rather than the other, as did the thematic materials and disjunctive rule-forms in the present experiments; or one could run a task where the implication interpretation of the conditional (and the inclusive disjunctive) were strongly specified: matching bias should not change relative to a version where implication and exclusion can be assumed by the

subject. The interpretations should though: the asymmetrical truth-tables associated with implication and inclusion should show at least some of the subjects that avoiding the negatives will not work - there should be more strictly logical behaviour on the negated rules as a result.

Re-interpretation should also be applied to the Selection task, as matching bias was. Four of the five published Selection tasks using negated rules are reported here (Experiments 1 and 2), and as they do not diverge in their findings this application can proceed with reference to the current data. Application of the re-interpretational theory to the Selection task is more problematical than it is to alternative truth-table tasks; bearing in mind the often-reported lack of transfer between truth-table and Selection task behaviour, beginning with Wason (1968), it is not easy to make strong predictions as to which interpretations subjects will apply in the latter task. This is not to duck the issue: if defective truth-tables are used in the treatment of conditionals, as in Experiments 7 and 8, they should also be evident in the Selection tasks of Experiments 1 and 2, since the same linguistic materials were used. We should therefore expect fewer cards exemplifying the $\overline{pq}$ case to be selected than cards exemplifying other cases; assuming that the $\overline{p}$ and $\overline{q}$ cards will be taken as exemplars of the 'defective' instance, we would therefore expect lower selection frequencies of these cards relative to the p and q cards. This is also a prediction of matching bias. If we apply strong predictions from the results of Experiments 7 and 8, we may also expect the selection patterns to the AN and NA rules to follow those normally found to an unnegated disjunctive, and the pattern on the NN rule to

follow the one on the AA. Failing this, and bearing in mind the possibility of the effects of the experimental situation, a weaker prediction would be that there should be some evidence on the negated rules of some kind of re-interpretation of those rules, competing with an appreciation of their true conditional structure, as was found in the truth-table tasks. There should also be clearer, and possibly different, response patterns under thematic materials. Let us see what happens.

For the convenience of the reader, the total response frequencies from Experiments 1 and 2 are collated in Table 31, and presented in terms of the matching values of the cards, i.e. assuming the avoidance of the negatives. Abstract and thematic data are presented separately; it will be recalled that both groups produced almost identical results, so the following remarks are mainly addressed to both together. Firstly, there were indeed fewer $\overline{p}$ and $\overline{q}$ cards selected, as the tests for matching bias found. Looking at the modal selection frequencies, however, it is immediately apparent that the split between AA and NN rules on the one hand and AN and NA rules on the other, found in the truth-table tasks, has not generalised to the Selection task. Rather, there is an equally clear divide here between the AA and AN rules and the NA and NN rules. Responses to the AA rules are well in line with previous findings (for the abstract task), where selections of the p and q cards are almost always the most frequent responses. The patterns on the AN rules here are almost identical: subjects seem to be ignoring the negative in the consequent. There is a vestige of evidence that some subjects were adhering to the logical structure both of the task and the rules. On both rules, almost all subjects

TABLE 31 Summary selection frequencies from Experiments 1 and 2, ordered by matching values. Notation as in Tables 6, 7, and 8. N = 48, in each group.

| | Thematic | | | | Abstract | | | |
|---|---|---|---|---|---|---|---|---|
| | p | $\bar{p}$ | q | $\bar{q}$ | p | $\bar{p}$ | q | $\bar{q}$ |
| AA | 44 | 6 | 24 | 14 | 43 | 7 | 31 | 14 |
| AN | 43 | 6 | 34 | 5 | 47 | 3 | 36 | 13 |
| NA | 17 | 35 | 41 | 24 | 14 | 39 | 36 | 19 |
| NN | 22 | 32 | 32 | 18 | 13 | 37 | 30 | 22 |

selected the (correct) p card; slightly more subjects selected the q card on the AN rule than on the AA, and more selected the $\bar{q}$ card on the AA than on the AN, where these cards form the other potential falsifier. The apparent difference here between the groups on logical behaviour did not reach significance in the individual experiments.

The response frequencies on the NA and NN rules were quite different. Here there were fewer p cards selected ($p < .001$, Sign test), more $\bar{q}$ cards ($p < .01$, Sign test), and the same number of q cards, compared with the AA and AN rules. Evans' explanation for most of these effects is quite straightforward. Logic and matching interact; if it is assumed that the logical component predominates on the antecedent, a high frequency of $\bar{p}$ selections would be expected on rules with negated antecedents, where it forms the true-antecedent card. There should still be a large proportion of p and q selections due to matching. It is not quite so clear why there should be more $\bar{q}$ selections, as the negation and mismatching of consequent items is mixed in the same way as in the other two rules. This suggests that there were, in fact, different interpretations applied to the NA and NN rules. Perhaps it is these which were being interpreted as disjunctives in this situation. Wason & Johnson-Laird (1972) did maintain, after all, that it was rules with negated antecedents which would tend to be treated in this way, and this might include both NA and NN forms. To establish this here, we need some idea of what the selection patterns would be when the task is presented in disjunctive form, and luckily there are three studies in the literature where this has been done. All three use the NA disjunctive, for its logical correspondence with the unnegated

conditional, but one (Wason & Johnson-Laird, 1969) also used an AA rule. With the NA rule, results from the three studies could hardly be more conflicting. Wason & Johnson-Laird, using abstract materials, presented the task in an unusual format, giving eight selection cards (i.e. two of each normal card, with both alternative values of the other rule component on their reverses) and the instruction that there were four which needed to be selected. They found a significant majority of correct selection combinations. Legrenzi (1970) however, using abstract materials in a strong exclusive sentence, found that 77% of his subjects chose the matching combination, and only 10% the fully insightful combination for exclusive disjunction - all four cards. Van Duyne (1974), using both abstract and thematic rules, found that under both types of materials responding was more or less random. The NA disjunctive is plainly up to its old tricks here. Wason & Johnson-Laird's experiment using the AA rule is all we have to go on for an indication of what we might expect of a simple disjunctive treatment of the Selection task, although, in view of their rather unexpected findings with the NA rule, we should perhaps be cautious. They assumed an inclusive reading of the rule and suspected in advance that the AA form would be "too easy", a suspicion borne out in the data, where a 75% correct response rate is recorded. The correct response for an inclusive disjunctive selection task is to select the $\overline{pq}$ combination, since only these could bear the one falsifying contingency ($\overline{pq}$). Evidently the ease of this task is in subjects' apprehension of the formal structure, both of the task and rules, since under a defective classification the $\overline{pq}$ case would be irrelevant (and the unnegated inclusive unfalsifiable). Should the

selections under an exclusive reading also entail a rejection of the defective table? We can only look to the data. If subjects were interpreting the rules, and treating the cards, as if the task involved defective exclusion, p and q selections should be in the majority. However, a reading of the rules as formal exclusion would lead to a tendency to select all four cards. What we have is such a tendency - there were more cards selected under the NA and NN rules than under AA and AN ($p < .001$, Sign test, 2-tailed) - with $\bar{p}$ and q the modal selections. This seems to indicate a competition between a formal exclusive and conditional treatment of these rules, with a tendency to read both as NA.

It is apparent that re-interpretation can only be spread very thinly over the Selection task. However, the lack of fit between the treatments seen in the truth-table tasks and the selections in the Selection tasks is not as serious as it seems. Obviously, a theory is stronger if it makes strong predictions and has them upheld; in the present account, one can go no further than to say that there are suggestions that defective truth-tables are applied to re-interpretations of the rules in the Selection task, without making exact a priori specifications as to what these interpretations are. A divergence between Selection and truth-table task performance is nothing new: it was a problem in Wason's (1968) original paper. In a recent article, Wason & Brooks (in press) point to a similar phenomenon in another fiendish logical task, the THOG problem, which is structured on the logic of exclusive disjunction: subjects seemed to understand the logic, but failed to apply it, much as they fail in the Selection task. These tasks clearly have an element of difficulty in them which defeats most of those who attempt them, and they clearly ask

different questions of subjects than do the 'underlying' truth-table tasks. It is not enough to interpret the rules, one must also apply that interpretation to half-concealed material. In both Selection and THOG tasks subjects seem tied to the perceptual elements of the materials; in the Selection task they may indeed focus on the cards named in the rules (Johnson-Laird & Wason, 1970a; Evans & Lynch, 1973), just as they seem unable to detach themselves from the visual attributes of the THOG stimuli. It is at the application stage that most people trip up, and in the present experiments, in the absence of any helpful circumstances, it was also sufficient to wipe out all but the merest suggestions of performance differences due to materials. Matching as focussing on named cards might profitably be retained in an account of Selection task performance; as such, it is unique to this task, rather than an extension of any truth-table evaluation.

The Selection task data do not allow a strengthening of the position of the re-interpretational theory relative to matching bias; Evans was able to make precise predictions on the basis of his findings on the truth-table task and demonstrate them on the Selection task (Evans & Lynch, 1973). The present theory was not. However, against that one can align the two signal advantages of re-interpretation in its explanation of truth-table data: it can account for the materials effect observed in Experiments 7 and 8, and it accounts for disjunctive performance, always the Achilles heel of matching bias (van Duyne, 1973; Evans, 1975; Evans & Newstead, in press). It can also go some way towards an account of the context effects suggested for the truth-table task and for which there is much evidence in the Selection task, as we saw in Chapter 5.

Implications: formal competence theories

To explain performance in reasoning tasks in terms of the recasting of rules and the application of truth-tables should not be mistaken for a retreat to the Henleian position, of presuming a basic formal ability obscured by idiosyncratic interpretations of the tasks. It may well be that there is a certain level of general 'logical' ability, inasmuch as people act consistently in structured situations. However, one cannot go from this to assert, or even describe, a 'natural logic' over which language casts its confusion. The performance of the Thematic subjects in Experiments 7 and 8 may give the appearance of rationality, if we assume that the avoidance of negation and the adoption of classifications which minimised its effects on the logical outcomes of the choices were interlinked. If they were just ignoring negatives and just using equivalence and exclusion strategies because both represent an easy way out, then rationality recedes. Until the problem is presented in a clearly defined, unambiguous logical structure in which negative-avoiding is not so excusable, the issue must remain unresolved. Certainly, the reality of formal competence is left in doubt when one surveys the content effects in the truth-table task, the context effects in the Selection task, and the lack of transfer between the simple truth-table task and the more difficult Selection task even when the same lexical materials are used in both.

This is in sharp contrast to those widely influential theories which are based on the idea of a formal logical competence which is tapped, or not, by reasoning tasks. By far the most influential approach of this kind is that of the Piagetian school, in its theory of formal operations. Such an approach must find

itself in opposition to the foregoing account of performance, which is in purely psychological terms and therefore essentially non-logical. Formalistic approaches cannot be ignored, especially when they emanate from illustrious sources, and so in the next section the Piagetian theory of formal operations will be evaluated in the light both of the present research and related literature. This will involve some reviewing of the theory, and an examination of its internal and external implications.

When confronted with the spectacle, in laboratories and outside, of people deliberating about problems and producing more or less sensible solutions, it is tempting to presume that there are some laws underlying their behaviour, and that these laws may form a structural system by which human rationality can be described. The formal calculus of propositional logic has been cast in this role in the past, but this idea, it has been noted, does not retain much currency among logicians or psychologists today. Rather, there have been repeated attempts at deriving systems of 'psychologic' wherein logic and psychology might meet and marry, and the system which has received the most concentrated attention has been that of Piaget and his followers. This attention has, outside the centre in Geneva where most Piagetian research is pursued, been largely critical, but this in itself is a measure of the impact and influence of Piaget's theories of cognitive competence and growth.

'Formal Operations' constitute the fourth and until recently the final stage of intellectual ontogeny which Piaget describes, the stage of development wherein the intellect reaches its final equilibrium. In the previous stages children are said to progress from a state of simple though organised activity,

through levels of perceptually dominated play and imitation, to the ability to classify and order objects in the real world and perform symbolic functions on these operations. These stages account for intellectual development up to the age of 11 or 12, whereupon a profound change in thinking takes place: whereas it has previously been tied to the real world, to the concrete objects before the senses, adolescent thought can now divorce itself from actuality. Where they once viewed the possible as an extrapolation of the real, a person in the stage of formal operations now views the real as a subset of the possible, he is now able "to reason about a proposition considered as a hypothesis independently of the truth of its content" (Beth & Piaget, 1966). Hence formal operations: reasoning on the logical form of the argument rather than, and apart from, its constituent material.

Like many other theories, Piagetian formal operations can take a strong form and a weak form, explanatory and predictive power diminishing sharply from the former to the latter. The weakened form has arisen through a series of retractions and modifications in answer to data gathered after the initial exposition.

Firstly then a description (taken mostly from Flavell, 1963; Piaget, 1957; and Beth & Piaget, 1966) of the strong form of the theory, since this is how it was first described, and how it may most usefully be applied to the present findings.

In the preceding developmental stage of concrete operations, the individual's cognitive task was primarily, as the name implies, to organise what was actually present, extrapolating the actual to the possible as the need arose; properties of objects could only be considered one by one, and the child could only perform operations on these properties one by one. The innovation

of the move into formal operations is provided by the ability to carry out a whole set of operations, to combine the separate operations used before into a coherent system of analysis. This is the basic strategy behind the new adolescent reasoning: not only can the reasoner think in terms of propositions about propositions, he can also do this in the form of a thorough analysis of all the potential variables in a problem. He subjects the problem to a combinatorial analysis, considering all facets of the possible to cross-check the actual. Piaget's description of how this applies to the consideration of a particular problem will illustrate how the individual uses this exhaustive analysis.

Taking the example of a problem of causality, the reasoner will ask himself two kinds of question. Firstly, he will ascertain whether fact x implies fact y, and in doing this he will cast the proposition as an implication ('if p then q') and look for the falsifying contingency, $p\bar{q}$. His second question will be whether it is fact x implying fact y, or whether y implies x, and he will test for this by checking for the absence of the falsifying case for this expression, $q\bar{p}$. Thus two of the operations which were carried out separately in the earlier concrete stage are combined - negation (N) and reciprocity (R). Together with two other operations, identity (I) and correlativity (C), these make up a logical 'four-group' of operations, INRC, a combinatorial system which the subject will use on any problem of this kind (Beth & Piaget, 1966).

There are immediate empirical problems with this formulation of adolescent (and presumably adult) reasoning. Piaget's writing on formal operations has been criticised by experimental psychologists for the less than perfect correspondence between the

theory and the evidence he himself adduces to support it, particularly in the book by Inhelder & Piaget (1958), which is its major exposition (e.g. Flavell, 1963; Lunzer, 1973). Detailed criticism of Inhelder & Piaget (1958) comes from Bynum, Thomas & Weitz (1972), who draw attention both to the small sample of behaviour reported and to imprecision and omission in the analysis. There are sixteen possible binary operations in an exhaustive application of the INRC logical group: Bynum et al. found evidence in Inhelder & Piaget (1958) for only eight of them being used by a subject. Six of the eight missing operations were found to have no equivalent natural-language expression, and could in fact be more simply described by alternative truth-functional operations, which leaves Bynum et al. wondering whether people ever use the complete INRC group. There is also an obvious correspondence between the logical behaviour described by Piaget and the structure of the tasks used in the present research and most of the studies reviewed in Part 1. There is a snag though: Piaget's description reads, as Wason has observed, like an accurate account of what the subjects are required but fail to do (although, strictly speaking, Beth & Piaget's account is of behaviour in a causal situation, and the tasks used here are nominally non-causal; see Wason & Johnson-Laird, 1972, ch. 14). In the case of the Selection task, a statement of implication is provided as part of the problem, so the subjects do not have to formulate it themselves, all they have to do is seek out the $p\bar{q}$ case. In the case of the truth-table tasks they do not even have to do this - they only have to recognise the cases as they appear. In both tasks, the use of abstract materials should remove the impediments to content-independence. In addition

we have also seen how it is possible to observe content effects in both problems. We therefore have two immediate points of departure from formal operations in tasks which appear eminently suitable for their application: not only do the solution patterns of highly intelligent young adults not accord with the logical behaviour explicitly required of 12 - 15 year-olds, but reasoning patterns also show a dependence on content and context. The model of reasoning outlined in the previous section, where probably learned solution strategies are pragmatically linked to the constituents of the problem in question, is fundamentally incompatible with the Piagetian scheme. Piaget is unequivocal in his opposition to the notion that 'logical' thinking is explainable in terms of experience, preferring to regard formal operations as "forms of equilibrium attained by thought activity" (Piaget, 1957). It is not easy to arrive at a precise grasp of what Piaget means when he talks of equilibrium in this way, especially when trying to discover how an individual might come to achieve it. Piaget's accounts of the transition from non-equilibrium to equilibrium states does not allow a clear conception of this acquisition or maturation process (Flavell, 1963).

The results of the research reported here and in the Review do not reconcile themselves easily with the requirements of formal operations, so we are left with two possibilities. The first is that the subjects in these experiments regressed to an earlier stage of development (cf. Wason, 1969a). This is implausible, not only because those subjects were not adolescents but intelligent young adults, but also because the tasks, especially the truth-table tasks, were not seen by the subjects as difficult - for instance,

the average verification times in Experiments 7 and 8 in this series were between three and fourteen seconds. The second possibility is that the theory of formal operations should be modified, or abandoned, and this leads both to the modifications arising from Geneva, and to an evaluation of alternative approaches.

Piagetian theory, as originally constituted and as later reworked, offers itself three possible resolutions in the face of seemingly contrary evidence. In the first place there is the concept of horizontal decalege, which states that while it may be possible to characterise an individual as being at a certain developmental stage and therefore possessing certain cognitive structures, the individual will not necessarily perform according to those structures on all tasks. As Flavell puts it, tasks differ in the extent to which they resist and inhibit the application of given structures. This makes the theory practically untestable. (Wason & Johnson-Laird, 1972; Ennis, 1975; Smedslund, 1977). Any positive result can be regarded as an example of the structures operating, and any negative result as being due to decalage. In any case, how would inhibition and resistance be assessed before testing? Leaving this weakness aside, there is the second possibility, that formal operations represent an ideal capability which subjects ordinarily never attain - in other words that the theory is one of competence rather than performance. This idea has much the same effect as decalage regarding testability, incurring as it does the difficulty which all models of 'pure' cognitive competence must face: the ideal conditions under which the true ability would emerge are elusive to the point of non-existence, and it is difficult to take prior account of imperfections in the situation. A final modification of the theory comes from

Piaget himself (Piaget, 1972): that formal operations are only observable within each individual's "area of specialisation". Being asked to reason about things of which he has no knowledge would hinder the subject's formal reasoning - lawyers would not be very good at reasoning about the theory of relativity, and in the same way physicists might not be too efficient at following the logic of the code of civil rights. Piaget seems to be positing a fifth stage of development here, a stage of specialisation, implying that the peak of reasoning ability is reached by the age of 15.

In this final modification of the theory of formal operations, Piaget is admitting, though he does not like to, some relation between form and content in reasoning. He does not go so far as, say, Wason & Johnson-Laird (1972), who give this interdependence some prominence. He seems to be saying that the extraction of logical form from an argument is most readily accomplished when dealing with familiar material. To this extent, he is specifying a context effect as much as a content effect, though the dividing line between the two, and between this and Wason & Johnson-Laird's content effect, is a thin one. In admitting the play of content and context effects, there are implications both for the strong form of the theory and for the present psychological approach. Firstly, it is plain that the strong form of the theory of formal operations cannot account for observed reasoning performance. Not only do people not do the same things, logical or otherwise, on the same arguments in different situations, they also do not follow formal logic in even very simple logical tasks such as the truth-table task, where subjects do not even have to operate on their appraisals. Piaget's admission of the role of content does bring his approach closer to the present one, but not by very

much, because his argument still rests on the extraction of formal structure. The present theory would rather say that familiar material would be more likely than unfamiliar material to invoke the learned solution strategies which have been found to be appropriate to that material in previous experience.

It has also been shown here that subjects will take the opportunity to indulge in some cognitive shortcutting, in their avoidance of negation and their preference for equivalence and exclusion over the more complicated relations of implication and inclusion. Perhaps this behaviour also gives some indications of a developmental angle on the present theory: that it is the learning of increasingly complex arguments, rather than the emergence of increasingly complex mental structures, which characterises cognitive development. It is not the wildest speculation to assert that human information-processing capacity expands through childhood to adulthood, or that this will be related to the complexity of arguments which can be handled at a given age. Piaget may be perfectly correct in maintaining that this general ability develops in discrete stages, but there is no need, under the present formulation, to go from this to the assertion of the development of specified logical structures. We might expect from this that deductive ability would be related to intelligence, assuming that intelligence and the handling of complex material are themselves related. In fact, there is evidence for such an idea, for instance Lunzer (1973) found that in tasks requiring complex inferences, performance was more closely related to IQ than to chronological or mental age. Below a certain age, children could not handle the problems at all, which indicates that

development of the abilities required may indeed proceed by discrete stages. On the more basic developmental point, Taplin, Staudenmayer & Taddonio (1974), studying children aged between nine and seventeen on problems of logical inference, found that the proportion of subjects in each age group responding consistently to some kind of truth function did not vary. Rather, it was the complexity of these functions which varied with age (the last point is equivocal owing to Taplin's dubious method of deriving truth-tables from inferences; see Chapter 3. This would not affect consistency ratings though). Similarly, Sinclair (cited by Piaget, 1970) found a correlation between children's logical performance and the complexity of their language output. This could mean that the logical tests were really testing lingusitic ability or, more likely, that both abilities are causally linked to general information-processing capacity.

Modern theories are still adopting Piaget's basic assumption that one can extrapolate, from some observed 'logical' behaviour, the possession by an individual or group of the complete logical arsenal. This assumption is a feature of the writings of those who attempt to define the 'natural logic' of certain linguistic expressions (e.g. Ennis, 1976; Braine, 1978). Braine has been criticised (Grandy, 1979) for presuming that 'If then' generally takes on the cognitive connotations of implication. Certainly, the case for this usage is, to put it mildly, not proven, but the critic himself also misses the point, in also arguing for acceptance of an alternative meaning of 'If then'. The mistake is to search for the one true meaning of the conditional, when it is perfectly possible for it to take a variety of 'meanings',

or 'treatments, to use the terminology of the present discussion, according to the requirements of the situation as the reasoner views them. It is possible to construct illustrative sentences for almost any logical connection (e.g. Ennis, 1976; Rips & Marcus, 1977; Braine, 1978, 1979), but this is not to say that, given a certain sentence, people's treatments of it will or should always follow the logic of it. Our striped tigers provided an example of this, and Ennis (1976) provides another. He maintains that people would be in trouble if they could not handle the logic of implication, but the sentence he uses as an illustration is revealing: 'If someone was a Soviet Communist at the time of the Vietnam war, then that person was opposed to what the United States was doing there'. This clearly does not imply its converse - or does it? It is not too difficult to imagine contexts in which, for the people concerned an equivalence treatment of this sentence might be ligitimate, expedient, or even neccessary. The logician cannot specify the pragmatic legitimacy of such a sentence for all individuals in all situations, much as he would like to.

## Conclusions

The view of the reasoner taken in the present discussion is, to borrow a phrase from personality theory, constructive rather than reactive (Pervin, 1975). In confronting a reasoning problem people do not simply roll, like Legrenzi's ball-bearings, down the rutted tracks of logic or matching, but actively rework the task material into forms which they can deal with, using treatments that have worked before. To the extent that they do not do this, as in the large proportion of classifications in the truth-table

Experiments 7 and 8 where no systematic alternative treatments seem to be applicable, they may be said to be committing errors. Whether such errors stem from simple bad logic, or the inaccessibility of the material to the application of learned solution strategies, is not a question the present discussion seeks to answer (Morley-Bunker, in press, argues vigorously for the latter explanation).

This view of reasoning also has something to say about the methods of reasoning research. Studies of reasoning have taken two principal forms: the Piagetian concrete approach and the linguistic/abstract approach adopted in the present research. The types of tasks used in either are fundamentally different. The typical Piagetian task uses a concrete, manipulative experiment such as is found in science classes at school, where the subject, usually a school-child, has to rearrange some object or objects to establish a physical relation, for instance, or a chemical proof. Such a task is probably well within most subjects' area of specialisation, at least at the time of testing. The only linguistic component in these tasks is in the running commentaries which Piagetian research usually requires of its subjects; language is certainly not the basis of the task itself. It is the basis of the tasks used in the present experiments however, and in most of the related studies cited in the previous chapters. One wonders whether the two methodologies are studying the same thing. In a Piagetian task the subject manipulates concrete material (not to be confused with thematic materials) in his own time, extracting what he will, or can, in terms of cognitive relations from it. Language does not proceed at such a leisured pace. The comprehension and evaluation of a sentence is literally a split-second affair and may be subject to change, to adapt to new information and circumstances. To a

strict Piagetian, and his critics, the question is vacuous: both tasks have logical structures and therefore must reflect a certain degree of logical competence, depending on the age of the subject. To a psychologist, rather than a logicist, the question is of no little importance. If subjects perform differently on different linguistic logical tasks, as has been found with the inference, truth-table, and Selection tasks, what might be the difference between two distinct classes of tasks?

The question becomes more urgent still when bearing in mind the Selection task and the formidable problems this presents to its victims and those who have to explain their behaviour. The Selection task is not a wholly lingustic task, it requires subjects to use language to operate on material which may or may not be itself linguistic. That it is, besides this, a wholly artificial task is irrelevant. Just because it is not an analogue of some real-life situation does not mean it has no relevance to real-life abilities. If it were totally alien, subjects would simply throw up their hands, make their excuses, and leave, or respond at random. They do not do this. Rather, it is a novel problem which people think they can solve, presumably by applying the abilities with which they were armed on entering the laboratory. The source of its difficulty, its divergence from, say, the truth-table task, lies in its demand that the subject engage in some reasoning (or meta-reasoning), something which the truth-table task does not. The truth-table task is, as the preceding discussion has indicated, more of an exhaustive psycholinguistic instrument for semantic analysis. It asks that subjects make the treatments they would give to conditional or disjunctive sentences explicit, by recognising and classifying instances from the universe of relevant

information from which that sentence might be drawn. This classification is a matter of the reading of the sentence and the consideration of the instance in parallel; the patterns of classification of sentences will vary according to the content of the sentences and the context in which they are encountered.

The question of context invites circularity into the argument. However, an infinity of possible contexts does not imply an infinity of context effects. It is possible to envisage some descriptive research which should eventually delineate a finite number of context effects and the ways in which they might interact with the treatment of sentences. To twist another phrase, this time one of Wason's, we need to spend some time investigating the contexts of plausible reasoning. Some progress has already been made: Rips & Marcus (1977) and Legrenzi (1970) have provided good information on the relation of causal and correlational contexts to equivalence classifications, which enabled an explanation of the treatment of the thematic materials used in the present study. There is also strong evidence of context effects in the Selection task, and the beginnings of a descriptive classification of these and their interaction with content in Chapter 5. It should be possible to manipulate these factors and predict and test for their effects in experiments.

This thesis started off by bemoaning formalistic psychology and arguing some consideration of content and context effects and response biases, and has ended by turning back on itself to some extent by reasserting, albeit tentatively, a degree of rationality for the subjects in reasoning experiments. In urging a consideration

of the pragmatic factors in reasoning performance, the overall view of human deductive ability must be a moderate one, its ground lying somewhere between the militant wings represented by Piaget's view of universal logical competence and Evans' view of substantial non-logical performance. Formal systems cannot describe the processes of reasoning, they can only, as Grandy (1979) points out, list the known alternatives and their formal differences; but neither may it fair to the subjects to characterise them as passive perceptually biased responders. The rationality which people exhibit depends on the strategies they have learned to apply, the materials they have to work with, and the situation which brings them all together.

REFERENCES


ALLPORT, F. H. (1920) The influence of the group upon association and thought. Cited by Zajonc, R. B. (1965), op. cit.

BEGG, I. & DENNY, J. P. (1969) Empirical reconciliation of atmosphere and conversion interpretation of syllogistic reasoning. Journal of Experimental Psychology, 81, 351-354.

BETH, E. W. & PIAGET, J. (1966) Mathematical epistemology and psychology. Dordrecht: Reidel.

BRACEWELL, R. J. (1974) Interpretation factors in the four card selection task. Paper presented at the Trento conference on the Selection task.

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.

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

BRAINE, M. D. S. (1979) If-then and strict implication: a response to Grandy's note. Psychological Review, 86, 154-156.

BREE, D. S. & COPPENS, G. (1976) The difficulty of an implication task. British Journal of Psychology, 67, 579-586.

BYNUM, T., THOMAS, J. A. & WEITZ, L. J. (1972) Truth-functional logic in formal operational thinking: Inhelder & Piaget's evidence. Developmental Psychology, 7, 129-132.

CARPENTER, P. A. & JUST, M. A. (1975) Sentence comprehension: a psycholinguistic processing model of verification. Psychological Review, 82, 45-73.

CERASO, J. & PROVITERA, A. (1971) Sources of error in syllogistic reasoning. Cognitive Psychology, 2, 400-410.

CHAPMAN, L. J. & CHAPMAN, J. P. (1959) Atmosphere effect re-examined. Journal of Experimental Psychology, 58, 220-226.

CLARK, H. H. & CHASE, W. G. (1972) On the processes of comparing sentences against pictures. Cognitive Psychology, 3, 472-517.

COPE, D. E. (1979) Reasoning with conditionals: the effects of a binary restriction. British Journal of Psychology, 70, 121-126.

CORNISH, E. R. (1971) Pragmatic aspects of negation in sentence evaluation and completion tasks. British Journal of Psychology, 62, 505-511.

CORNISH, E. R. & WASON, P. C. (1970) The recall of affirmative and negative sentences in an incidental learning task. Quarterly Journal of Experimental Psychology, 22, 109-114.

ENNIS, R. H. (1975) Children's ability to handle Piaget's propositional logic: a conceptual critique. Review of Educational Research, 45, 1-41.

ENNIS, R. H. (1976) An alternative to Piaget's conception of logical competence. Child Development, 17, 903-919.

ERICKSON, J. R. & JONES, M. R. (1978) Thinking. Annual Review of Psychology, 29, 61-90.

EVANS, J. St. B. T. (1972a) Reasoning with negatives, British Journal of Psychology, 63, 213-219.

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

EVANS, J. St. B. T. (1972c) On the problems of interpreting reasoning data: logical and non-logical approaches. Cognition, 1, 373-384.

EVANS, J. St. B. T. (1975) On interpreting reasoning data: a reply to van Duyne. Cognition, 3, 387-390.

EVANS, J. St. B. T. (1977a) Linguistic factors in reasoning. Quarterly Journal of Experimental Psychology, 29, 297-306.

EVANS, J. St. B. T. (1977b) Toward a statistical theory of reasoning. Quarterly Journal of Experimental Psychology, 29, 621-635.

EVANS, J. St. B. T. (1978) The psychology of deductive reasoning: logic. In Burton, A. & Radford, J. (Eds.) Thinking in perspective. London: Methuen.

EVANS, J. St. B. T. & LYNCH, J. S. (1973) Matching bias in the selection task. British Journal of Psychology, 64, 391-397.

EVANS, J. St. B. T. & NEWSTEAD, S. E. (1977) Language and reasoning: a study of temporal factors. Cognition, 5, 265-283.

EVANS, J. St. B. T. & NEWSTEAD, S. E. (in press) A study of disjunctive reasoning. Psychological Research.

EVANS, J. St. B. T & WASON, P. C. (1976) Rationalisation in a reasoning task. British Journal of Psychology, 67, 479-486.

FESTINGER, L. (1957) A theory of cognitive dissonance. New York: Harper & Row.

FILE, S. E. & JEW, A. (1973) Syntax and the recall of instructions in a realistic situation. British Journal of Psychology, 64, 65-70.

FILLENBAUM, S. (1974a) 'Or': some uses. Journal of Experimental Psychology, 103, 913-921.

FILLENBAUM, S. (1974b) Pragmatic normalisation: further results for some conjunctive and disjunctive sentences. Journal of Experimental Psychology, 102, 574-578.

FLAVELL, J. H. (1963) The developmental psychology of Jean Piaget. London: Van Nostrand.

FRASE, L. T. (1966) Validity judgements of syllogisms in relation to two sets of terms. Journal of Educational Psychology, 57, 239-245.

FRASE, L. T. (1968) Effects of semantic incompatibility upon deductive reasoning. Psychonomic Science, 12, 64.

GEIS, M. & ZWICKY, A. M. (1971) On invited inferences. Linguistic Inquiry, 2, 561-566.

GILHOOLY, K. J. & FALCONER, W. A. (1974) Concrete and abstract terms and relations in testing a rule. Quarterly Journal of Experimental Psychology, 26, 355-359.

GOODWIN, R. Q. & WASON, P. C. (1972) Degrees of insight. British Journal of Psychology, 63, 205-212.

GRANDY, R. E. (1979) Inference and if-then. Psychological Review, 86, 152-153.

GREENE, J. M. (1970) Syntactic form and semantic function. Quarterly Journal of Experimental Psychology, 22, 14-27.

HENLE, M. (1962) On the relation between logic and thinking. Psychological Review, 69, 366-378. Reprinted in Wason, P. C. & Johnson-Laird, P. N. (Eds.) Thinking and reasoning. Harmondsworth: Penguine (1968)

INHELDER, B. & PIAGET, J. (1958) The growth of logical thinking. New York: Basic Books.

JANIS, I. L. & FRICK, F. (1943) The relationship between attitudes towards conclusions and errors in judging logical validity of syllogisms. Journal of Experimental Psychology, 33, 73-77

JOHNSON-LAIRD, P. N. (1975) Models of deduction. In Falmagne, R. J. (Ed.) Reasoning: representation and process. Hillsdale: Erlbaum.

JOHNSON-LAIRD, P. N., LEGRENZI, P. & LEGRENZI, M. S. (1972) Reasoning and a sense of reality. British Journal of Psychology, 63, 395-400.

JOHNSON-LAIRD, P. N. & TAGART, J. (1969) How implication is understood. American Journal of Psychology, 82, 367-373.

JOHNSON-LAIRD, P. N. & TRIDGELL, J. M. (1972) When negation is easier than affirmation. Quarterly Journal of Experimental Psychology, 24, 87-91.

JOHNSON-LAIRD, P. N. & WASON, P. C. (1970) A theoretical analysis of insight into a reasoning task. Cognitive Psychology, 1, 134-148.

JOHNSON-LAIRD, P. N. & WASON, P. C. (Eds.) (1977) Thinking: readings in cognitive science. London: Cambridge University Press.

LEGRENZI, P. (1970) Relations between language and reasoning about deductive rules. In FLORES D'ARCAIS, G. B. & LEVELT, W. J. M. (Eds.) Advances in Psycholinguistics. Amsterdam: North-Holland.

LUNZER, E. A. (1973) The development of formal reasoning: some recent experiments and their implications. In Frey, K. & Lang, M. (Eds.) Cognitive Processes and Science Instruction. Baltimore: Williams & Wilkins.

LUNZER, E. A. (1975) The development of advanced reasoning abilities. Italian Journal of Psychology, 2, 369-390.

LUNZER, E. A., HARRISON, C. & DAVEY, M. (1972) The four-card problem and the generality of formal reasoning. Quarterly Journal of Experimental Psychology, 24, 326-339.

MEHLER, J. (1963) Some effects of grammatical transformation on the recall of English sentences. Journal of Verbal Learning and Verbal Behaviour, 2, 346-351.

MORLEY-BUNKER, N. (in press) The presumption of logicality in reasoning. Acta Psychologica.

MORGAN, W. J. & MORGAN, A. B. (1953) Logical reasoning: with and without training. Journal of Applied Psychology, 37, 399-401.

MORGAN, J. J. & MORTON, J. T. (1944) The distortion of syllogistic reasoning produced by personal convictions. Journal of Social Psychology, 20, 39-59.

MOSHMAN, D. (1978) Some comments on Bree & Coppens' 'The difficulty of an implication task'. British Journal of Psychology, 69, 371-372.

PARIS, S. G. (1973) Comprehension of language connectives and propositional logical relationships. Journal of Experimental Child Psychology, 16, 278-291.

PERVIN, L. A. (1975) Personality: theory, assessment, and research. London: Wiley.

PIAGET, J. (1957) Logique et equilibre dans le comportement du sujet. Cited by Gruber, H. E. & Voneche, J. H. (1977). The Essential Piaget. London: Routledge & Kegan Paul.

PIAGET, J. (1970) Genetic epistemology. New York & London: Columbia University Press.

PIAGET, J. (1972) Intellectual evolution from adolescence to adulthood. Reprinted in JOHNSON-LAIRD, P. N. & WASON, P. C. (Eds.) (1977), op. cit.

POLLARD, P. & EVANS, J. St. B. T. (in press) The effects of prior beliefs in reasoning: an associational interpretation. British Journal of Psychology.

REVLIN, R. & LEIRER, V. O. (1978) The effects of personal biases on syllogistic reasoning: rational decisions from personalised experimentation. In Revlin, R. & Mayer, J. (Eds.) Human Reasoning. Washington D. C.: Winston.

RIPS, L. J. & MARCUS, S. (1977) Suppositions and the analysis of conditional sentences. In JUST, J. A. & CARPENTER, P. A. EEds.) Cognitive Processes in Comprehension. Hillsdale: Erlbaum

ROBERGE, J. J. (1971a) An analysis of response patterns for conditional reasoning schemes. Psychonomic Science, 22, 338-339.

ROBERGE, J. J. (1971b) Some effects of negation on adults' conditional reasoning abilities. Psychological Reports, 29, 839-844.

ROBERGE, J. J. (1974) Effects of negation on adults' comprehension of fallacious conditional and disjunctive arguments. Journal of General Psychology, 91, 287-293.

ROBERGE, J. J. (1976a) Effects of negation on adults' disjunctive reasoning abilities. Journal of General Psychology, 94, 23-28.

ROBERGE, J. J. (1976b) Reasoning with exclusive disjunction arguments. Quarterly Journal of Experimental Psychology, 28, 419-427.

ROBERGE, J. J. (1977) Effects of content on inclusive disjunction reasoning. Quarterly Journal of Experimental Psychology, 29, 669-676.

ROBERGE, J. J. (1978) Linguistic and psychometric factors in propositional reasoning. Quarterly Journal of Experimental Psychology, 30, 705-716.

SELLS, S. B. (1936) The atmosphere effect: An experimental study of reasoning. Archives of Psychology, 29, 3-72.

SMALLEY, N. S. (1974) Evaluating a rule against possible instances. British Journal of Psychology, 65, 293-304.

SMEDSLUND, J. (1977) Piaget's psychology in practice. British Journal of Educational Psychology, 47, 1-6.

STAUDENMAYER, H. (1975) Understanding conditional reasoning with meaningful propositions. In FALMAGNE, R. J. (Ed.) Reasoning: representation and process. Hillsdale: Erlbaum.

STRAWSON, P. F. (1959) Introduction to logical theory. London: Methuen.

SUPPES, P. (1965) On the behavioural foundations of mathematical concepts. Cited by WASON, P. C. & JOHNSON-LAIRD, P. N. (1972) op. cit.

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.

TAPLIN, J. E., STAUDENMAYER, H. & TADDONIO, J. L. (1974) Developmental changes in conditional reasoning: linguistic or logical? Journal of Experimental Child Psychology, 17, 360-373.

TRABASSO, T., ROLLINS, H. & SHAUGHNESSY, E. (1971) Storage and verification stages in processing concepts. Cognitive Psychology, 2, 239-289.

VAN DUYNE, P. C. (1973) A short note on Evans' criticism of reasoning experiments and his matching bias hypothesis. Cognition, 2, 239-242.

VAN DUYNE, P. C. (1974) Realism and linguistic complexity in reasoning. British Journal of Psychology, 65, 59-67.

VAN DUYNE, P. C. (1976) Necessity and contingency in reasoning. Acta Psychologica, 40, 85-101.

WASON, P. C. (1959) The processing of positive and negative information. Quarterly Journal of Experimental Psychology, 11, 92-107.

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

WASON, P. C. (1964) The effect of self-contradiction on fallacious reasoning. Quarterly Journal of Experimental Psychology, 16, 30-34. Reprinted in WASON, P. C. & JOHNSON-LAIRD, P. N. (Eds.) (1968) Thinking and reasoning. Harmondsworth: Penguin.

WASON, P. C. (1965) The contexts of plausible denial. Journal of Verbal Learning and Verbal Behaviour, 4, 7-11.

WASON, P. C. (1966) Reasoning. In FOSS, B. M. (Ed.) New Horizons in Psychology, 1. Harmondsworth: Penguin.

WASON, P. C. (1968) Reasoning about a rule. Quarterly Journal of Experimental Psychology, 20, 273-281.

WASON, P. C. (1969a) Regression in reasoning? British Journal of Psychology, 60, 471-480.

WASON, P. C. (1969b) Structural simplicity and psychological complexity: some thoughts on a novel problem. Bulletin of the British Psychological Society, 22, 281-284.

WASON, P. C. (1972) In real life negatives are false. Logique et Analyse, 57-58, 17-38.

WASON, P. C. (1977) Self-contradictions. In JOHNSON-LAIRD, P. N. & WASON, P. C. (Eds.) (1977) op. cit.

WASON, P. C. & BROOKS, P. G. (in press). THOG: the anatomy of a problem. Psychological Research.

WASON, P. C. & EVANS, J. St. B. T. (1975) Dual processes in reasoning? Cognition, 3, 141-154.

WASON, P. C. & GOLDING, E. (1974) The language of inconsistency. British Journal of Psychology, 65, 537-546.

WASON, P. C. & JOHNSON-LAIRD, P. N. (1969) Proving a disjunctive rule. Quarterly Journal of Experimental Psychology, 21, 14-20.

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.

WASON, P. C. & JOHNSON-LAIRD, P. N. (1972) Psychology of reasoning: structure and content. London: Batsford.

WASON, P. C. & JONES, S. (1963) Negatives: denotation and connotation. British Journal of Psychology, 54, 299-307.

WASON, P. C. & SHAPIRO, D. A. (1971) Natural and contrived experience in a reasoning problem. Quarterly Journal of Experimental Psychology, 23, 63-71.

WILKINS, M. C. (1928) The effects of changed material on the ability to do formal syllogistic reasoning. Archives of Psychology, 16, No. 102.

WOODWORTH, R. S. & SELLS, S. B. (1935) An atmosphere effect in formal syllogistic reasoning. Journal of Experimental Psychology, 18, 451-460.

ZAJONC, R. B. (1965) Social facilitation. Science, 149, 269-274.

APPENDIX A

Example of a test sheet, as used in Experiments 1 and 2

Rule:

If I eat pork then I do not drink wine

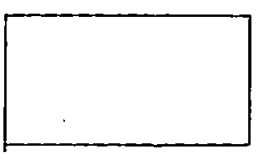
Please indicate which of the cards drawn below would need to be turned over to find out whether the rule has been obeyed or not. Please tick (✓) any of the cards you think would need to be turned over, and cross (X) any you think would not need to be turned over. Please don't leave any unmarked.

Turn over?

 .....

 .....

 .....

 .....

APPENDIX B

Analysis of variance of log. latency scores from Experiment 6. Significance levels for all repeated-measures factors are assessed using conservative degrees of freedom.
Factors are G (groups), B (blocks) and T (truth-table case).

| Source | SS | df | MS | F | Sig. |
|---|---|---|---|---|---|
| Between Ss | | | | | |
| G | 0.004 | 1 | .00440 | <1 | NS |
| Error | 19.323 | 30 | .64409 | | |
| Within Ss | | | | | |
| B | 5.359 | 5 | 1.07189 | 44.92 | p <.001 |
| BG | .074 | 5 | .01489 | <1 | NS |
| Error | 3.580 | 150 | .02386 | | |
| T | 5.514 | 3 | 1.83799 | 28.32 | p <.001 |
| TG | .275 | 3 | .09152 | 1.41 | NS |
| Error | 5.840 | 90 | .06489 | | |
| BT | .350 | 15 | .02334 | 1.14 | NS |
| BTG | .222 | 15 | .01482 | <1 | NS |
| Error | 9.181 | 450 | .02040 | | |

APPENDIX C

Analysis of variance of log. Comprehension times from Experiment 7. Conservative degrees of freedom used (see Appendix D).
Factors are G (groups), A (polarity of antecedent), and C (polarity of consequent). For CTs, truth-table case way a dummy factor which did not affect the analysis, and this factor is therefore not included in the table.

| Source | SS | df | MS | F | Sig. |
|---|---|---|---|---|---|
| Between Ss | | | | | |
| G | 4.2514 | 1 | 4.251 | 14.02 | p < 01 |
| Error | 6.6724 | 22 | .303 | | |
| Within Ss | | | | | |
| A | .7850 | 1 | .785 | 10.24 | p <.01 |
| AG | .7171 | 1 | .717 | 9.35 | p <.01 |
| Error | 1.6873 | 22 | .076 | | |
| C | .5627 | 1 | .563 | 17.92 | p <.001 |
| CG | .0177 | 1 | .018 | <1 | NS |
| Error | .6907 | 22 | .031 | | |
| AC | .0080 | 1 | .008 | <1 | NS |
| ACG | .0242 | 1 | .024 | <1 | NS |
| Error | .7569 | 22 | .034 | | |

APPENDIX D

Analysis of variance of log. Verification times from Experiment 7. Conservative degrees of freedom used (see Appendix B). Factors are G (groups), T (truth-table case), A (polarity of antecedent), C (polarity of consequent).

| Source | SS | df | MS | F | Sig. |
|---|---|---|---|---|---|
| Between Ss | | | | | |
| G | .0452 | 1 | .045 | <1 | NS |
| Error | 7.9221 | 22 | .360 | | |
| Within Ss | | | | | |
| T | 1.5175 | 3 | .506 | 8.47 | p <.01 |
| TG | .2652 | 3 | .088 | 1.48 | NS |
| Error | 3.9392 | 66 | .059 | | |
| A | 1.4186 | 1 | 1.419 | 29.96 | p < .001 |
| AG | .0008 | 1 | .001 | <1 | NS |
| Error | 1.0418 | 22 | .047 | | |
| C | 1.3904 | 1 | 1.390 | 14.59 | p <.001 |
| CG | .0011 | 1 | .001 | <1 | NS |
| Error | 2.0965 | 22 | .095 | | |
| TA | .2140 | 3 | .071 | 1.79 | NS |
| TAG | .0567 | 3 | .019 | <1 | NS |
| Error | 2.6264 | 66 | .039 | | |
| TC | .4822 | 3 | .161 | 6.18 | p <.01 |
| TCG | .0464 | 3 | .015 | <1 | NS |
| Error | 1.7166 | 66 | .026 | | |
| AC | .0001 | 1 | .000 | 0 | NS |
| ACG | .1065 | 1 | .106 | 1.99 | NS |
| Error | 1.1796 | 22 | .053 | | |
| ACT | .7927 | 3 | .264 | 8.05 | p <.01 |
| ACTG | .5953 | 3 | .198 | 6.05 | p <.01 |
| Error | 2.1659 | 66 | .032 | | |

APPENDIX E

Analysis of variance of log. Verification times from Experiment 7. Conservative degrees of freedom (see Appendix B).

Factors are G (groups), M (matching case), and R (rules)..

| Source | SS | df | MS | F | Sig. |
|---|---|---|---|---|---|
| Between Ss | | | | | |
| G | .0402 | 1 | .040 | <1 | NS |
| Error | 7.8943 | 22 | .358 | | |
| Within Ss | | | | | |
| M | .8951 | 3 | .298 | 9.47 | p <.01 |
| MG | .4808 | 3 | .160 | 5.09 | p <.01 |
| Error | 2.0789 | 66 | .031 | | |
| R | 2.8062 | 3 | .935 | 14.29 | p <.01 |
| RG | .1017 | 3 | .034 | <1 | NS |
| RM | 2.1269 | 9 | .236 | 5.59 | p <.05 |
| RMG | .4779 | 9 | .053 | 1.26 | NS |
| Error | 8.3685 | 198 | .042 | | |

APPENDIX F

Analysis of variance of log. Comprehension times from Experiment 8. Conservative degrees of freedom (see Appendix B).
Factors are G (groups), F (rule-forms), A (polarity of antecedent), C (polarity of consequent). Truth-table case was a dummy factor (see Appendix C).

| Source | SS | df | MS | F | Sig. |
|---|---|---|---|---|---|
| Between Ss | | | | | |
| G | .0373 | 1 | .037 | <1 | NS |
| Error | 51.1168 | 30 | 1.704 | | |
| Within Ss | | | | | |
| F | 3.3098 | 2 | 1.655 | 15.51 | p <.001 |
| FG | .5666 | 2 | .283 | 2.66 | NS |
| Error | 6.4001 | 60 | .106 | | |
| A | 4.8320 | 1 | 4.832 | 72.05 | p <.001 |
| AG | .1146 | 1 | .115 | 1.71 | NS |
| Error | 2.0119 | 30 | .067 | | |
| C | 1.7122 | 1 | 1.712 | 35.89 | p <.001 |
| CG | .0006 | 1 | .001 | <1 | NS |
| Error | 1.4312 | 30 | .047 | | |
| AF | .0854 | 2 | .043 | <1 | NS |
| AFG | .1566 | 2 | .078 | 1.41 | NS |
| Error | 3.3394 | 60 | .055 | | |
| CF | .1983 | 2 | .099 | 2.94 | NS |
| CFG | .1323 | 2 | .066 | 1.96 | NS |
| Error | 2.0244 | 60 | .033 | | |
| AC | .9402 | 1 | .940 | 16.28 | p <.001 |
| ACG | .0169 | 1 | .017 | <1 | NS |
| Error | 1.7324 | 30 | .057 | | |
| ACF | 1.5443 | 2 | .772 | 12.84 | p <.01 |
| ACFG | .1381 | 2 | .069 | 1.15 | NS |
| Error | 3.6073 | 60 | .060 | | |

APPENDIX G

Analysis of variance of log. Verification times from Experiment 8.

Conservative degrees of freedom (see Appendix B).

Factors are G (groups), F (rule-form), M (matching case), R (rules).

| Source | SS | df | MS | F | Sig. |
|---|---|---|---|---|---|
| Between Ss | | | | | |
| G | .2020 | 1 | .202 | <1 | NS |
| Error | 43.8947 | 30 | 1.463 | | |
| Within Ss | | | | | |
| F | 1.2776 | 2 | .639 | 5.54 | p <.01 |
| FG | .1914 | 2 | .096 | <1 | NS |
| Error | 6.9242 | 60 | .115 | | |
| M | 1.4884 | 3 | .496 | 5.29 | p <.01 |
| MG | 1.4381 | 3 | .479 | 5.11 | p <.01 |
| Error | 8.4353 | 90 | .093 | | |
| R | 5.2603 | 3 | 1.753 | 20.10 | p <.001 |
| RG | 1.0703 | 3 | .357 | 4.09 | p <.01 |
| Error | 7.8518 | 90 | .087 | | |
| FM | 4.3139 | 6 | .719 | 10.77 | p <.01 |
| FMG | .2134 | 6 | .036 | <1 | NS |
| Error | 12.0166 | 180 | .066 | | |
| FR | 2.4405 | 6 | .407 | 6.50 | p <.05 |
| FRG | .6084 | 6 | .101 | 1.62 | NS |
| Error | 11.2675 | 180 | .063 | | |
| MR | 5.2434 | 9 | .583 | 12.65 | p <.01 |
| MRG | .7261 | 9 | .081 | 1.75 | NS |
| Error | 12.4361 | 270 | .046 | | |
| FMR | 4.7245 | 18 | .262 | 4.71 | p <.05 |
| FMRG | 1.0859 | 18 | .060 | 1.08 | NS |
| Error | 30.0609 | 540 | .056 | | |