Chapter 3
"Differences in Experience": Proposed explanations for the elusivity of the content effect on the Wason selection task

A number of theories attempting to explain the elusive content effect on the Wason selection task have appeared in the literature. Most agree that thematic content enhances logical performance because thematic rules are familiar, whereas abstract rules are unfamiliar. The theories differ in their explanations of why familiarity enhances performance, and why the content effect is so "elusive."

None of these theories invoke the notion of a social contract, therefore none of them try to explain why social contract rules are the only thematic rules to consistently elicit robust content effects. Invoking the concept of a social contract turns the theoretical problem on its head: the phenomenon requiring explanation is not the content effect's elusivity, but, rather, its predictability.

3.1 Families of explanation

Before discussing the particular theories that have already been proposed, it is useful to consider what kinds of explanation are possible in general. Conceptually, there are at least five relatively distinct families of explanation:

1. There is no logic module. In solving the selection task, people use rules of inference appropriate to the domain suggested by the problem. These rules of inference may be different for different content domains.
a. The rules of inference are a product of "experience" structured only by information processing mechanisms that are innate, but domain general.

b. The rules of inference are innate, or else the product of "experience" structured by domain specific innate algorithms.

2. There is a logic module, but it is not necessary for everyday learning. It is activated only in higher level model building, for example, to answer questions within the framework of a well-established theory of what is true of a content domain. That is why performance is better with familiar materials.

3. There is a logic module, and it is necessary for learning. The content effect is due to differences in how well the propositions can be pushed through auxiliary mechanisms like short term stores or imagery buffers. Familiar terms and/or relations facilitate performance because they are concrete and therefore more easily manipulable or because they reduce "cognitive load".

4. There is no logic module, just the ability to recognize contradiction when one sees it.

   a. People can build mental models of the circumstances described in a problem; if they happen to build a model that contradicts the state of affairs asserted by the conditional, they will falsify. It is easier to build mental models of familiar propositions and relations.

   b. Actual experience with events that contradict the relation are stored in long-term memory. A familiar theme is more likely to cue contradictory associational pairing from long term memory, because such pairings are more likely to have been actually experienced.

5. Non-rational, domain-general heuristics having nothing to do with formal logic, or with an understanding of the relevance of counter-examples, account for the presence and variability of the content effect.

The hypothesis that humans have Darwinian algorithms for reasoning about social exchange is a "family 1-b" explanation. Each explanation proposed in the literature belongs to one of these five families of explanation.
3.2 Explanations proposed in the literature

A number of explanations have been put forth to explain content effects on the Wason selection task. Most of them involve a wedding of associationism and Tversky & Kahneman's (1973) "availability" heuristic.

Tversky & Kahneman were interested in how people judge probability. They noted that people typically do not make statistically sound probabilistic inferences, even when given information sufficient to do so.

Although people's probability judgments are not statistically sound, they are not random, either. To account for this, Tversky & Kahneman posited that people use mental short cuts -- "heuristics" -- in making probability judgments. They hypothesized that people judge the probability of two events co-occurring by the ease with which examples come to mind -- their "availability". They named this method the "availability heuristic."

For example, suppose you are told that 80% of college students in Cambridge attend Harvard and 20% attend MIT. A Cambridge college student was involved in a fight today. Your task is to guess which school this student attends. Five fights involving MIT students immediately spring to mind, but you have to search your memory long and hard to recall any fights involving Harvard students: the co-occurrence of "MIT" and "fight" is more available as a response. Even though Harvard students outnumber MIT students 4 to 1 in Cambridge, and even though you have no reliable data indicating that MIT students are more pugnacious than Harvard students, the availability heuristic
would lead you to judge that the fight today was more likely to have involved an MIT student than a Harvard student.

According to Tversky & Kahneman, ease of recall is a function of associative strength. Associative strength, they argue, is usually directly proportional to the frequency with which two events co-occur in an individual's experience. The availability heuristic is a useful rule of thumb because the ease with which associations can be brought to mind is usually correlated with their ecological frequency. It can lead to bias, however, when associative strength is determined by factors other than ecological frequency (like semantic distance or perceptual saliency).

Frequent events are familiar events. Abstract rules relating letters and numbers are unfamiliar. It occurred to a number of researchers that availability -- based on frequency-determined associative strength -- might play a key role in explaining why some familiar problems are more likely to elicit logical performance on the Wason selection task than abstract problems.

For Tversky & Kahneman, ecological frequency was only one of many determinants of availability. But because selection task theorists were trying to account for a content effect that they assumed was caused by familiarity, associationism plays a more central role in their adaptations of availability theory.

The "availability theories" of the selection task theorists come in a variety of forms, with some important theoretical differences. But common to all is the notion that the subject's actual past experiences create associational links between terms.
mentioned in the selection task. The more exposures a subject has had to, for example, the co-occurrence of P and Q, the stronger that association will be and the easier it will come to mind -- become "available" as a response. A subject is more likely to have actually experienced the co-occurrence of P and not-Q for a familiar rule, therefore familiar rules are more likely to elicit logically falsifying responses than unfamiliar rules. If all the terms in a task are unfamiliar, the only associational link available will be that created between P and Q by the conditional rule itself, because no previous link will exist among any of the terms. Thus 'P & Q' will be the most common response for unfamiliar rules.

Although it is rarely explicitly stated, these theorists seem to assume that associative links are created "the old-fashioned way", by domain general associative processes. Some refer directly to associationism (Pollard, 1982), whereas others refer more simply to the different amount of "experience" subjects may have had with different content domains (Griggs & Cox, 1982; Manktelow & Evans, 1979; Johnson-Laird, 1983; Wason, 1983). The presumption that learning occurs via some sort of "computational associationism" (Fodor, 1983) would account for their belief that the categorization of content domains along a familiar-unfamiliar dimension is the correct one, the one with causal import. Associationism is a process that makes unfamiliar content domains familiar -- regardless of the specific content of the domain it operates upon. Which content domains become familiar is determined by the amount of personal experience a particular individual has with the domains in question. The
selection task theorists rarely entertain the notion that regardless of familiarity, different content domains are processed by different, domain specific rules of inference. When they do, they seem to presume that the domain specific rules were learned via a domain general process.

The P card is almost universally chosen on Wason selection tasks, regardless of content. All theories that have been proposed in the literature concede that this is probably due to a rudimentary understanding of logic (or of contingency, in a looser, linguistic sense). Thus, the primary goal of these theories is to explain why familiar rules facilitate the selection of the not-Q card and inhibit the selection of the Q card, insofar as this happens. To be adequate, a theory must be able to answer three questions raised by the data reviewed in the previous chapter:

1. Why do familiar rules elicit more logical falsification than abstract rules?
2. Why do some familiar rules reliably elicit logical falsification whereas others do not?
3. Why do the same familiar rules sometimes elicit logical responses and sometimes not?

Differential Availability

In an article entitled "Human reasoning: Some possible effects of availability", Paul Pollard put forth what is to date the most precisely specified theory purporting to explain content effects on the Wason selection task (Pollard, 1982). It is a quite literal application of the associationist paradigm sketched above, in which pre-existing associative pairings of terms
mentioned in the selection task create a non-logical response bias (his theory is a "family 5" explanation). Whether a subject responds 'P & Q' or 'P & not-Q' is determined by the relative strength of these two associative links.* The dominant association wins, even if both are available. Thus, a subject will answer 'P & Q' if more instances of P - Q links come to mind than instances of P - not-Q links. For Pollard, associative strength is directly proportional to the number of exposures an individual has had to each pairing. Actual personal experience is the centerpiece of his availability theory.

For example, on a transportation problem where the rule is "If a person goes to Boston then he takes the subway" and the cards are "Boston" (P), "Arlington" (not-P), "subway" (Q), and "cab" (not-Q), a subject who had had more experiences of people taking the subway to Boston would choose "Boston" and "subway", that is, 'P & Q'. A subject who had had more experiences of people taking a cab to Boston would choose "Boston" and "cab", that is, 'P & not-Q', which is, by coincidence, the logically falsifying response. Note, however, that the procedure that generated this response is decidedly non-logical.

Pollard distinguishes between "realistic" content and content that is merely "thematic". Thematic content is not

* Pollard does not explicitly discuss why someone might choose 'P' alone on the selection task. However, in discussing other logical tasks he notes that availability might affect a conditional's perceived reversibility; "all dogs are animals" is clearly not the same as "all animals are dogs", whereas "all dogs bark" is not so clearly different from "all barking animals are dogs." From here he would have to argue that having understood that "If P then Q" does not imply "If Q then P" somehow prevents one from choosing the Q card. But since his theory is a nonlogical one, and nonreversibility is a logical consideration, it is not clear what that "somehow" would be.
"realistic" unless it cues actual experiences. If the subject has had no relevant experiences with the problem domain, no matter how "thematic" it is, the dominant association will be that created by the conditional rule itself. Hence, the subject will respond 'P & Q', just as if the problem's content were abstract. Pollard is a stickler for actual experience. For example, I can think of no relation that people have more experience with than that expressed by the food problem: "If I eat X then I drink Y". Most meals include both food and drink, and most people eat three such meals a day, every day of their lives. Moreover, it is quite common for certain foods and drinks to be consumed in conjunction with one another: orange juice with breakfast foods, coffee with dessert, wine with dinner entrees, mixed drinks with hors d'oeuvres.* Yet Pollard claims that the food problem did not elicit a content effect because subjects probably had not personally experienced some of the particular food-drink combinations used, such as, "If I eat haddock then I drink gin". (In some of my experiments I administered food problems using more usual content, and still found no effect, see Chapter 6.)

Because responses are determined by the actual, personal, idiosyncratic experiences of subjects, his theory can account for the fact that certain contents, like the transportation problem, sometimes elicit logical responses and sometimes do not:

* Note also that for meals, the most common eating plus drinking experiences, it is the type of food eaten that determines what drink is served, not vice versa. Like the food problem, the relation for meals is "If I eat X, then I drink Y", not "If I drink Y then I eat X).
...the extent of bias toward one mode of transport would be expected to vary from study to study and, to some extent, from subject to subject, depending on such factors as geographical location, income level of the subjects and the appearance of the experimenter himself (subjects, for instance, may well have experience of professors, but not of postgraduate students, reporting travel by plane). (pp. 80-81)

Unfortunately, for the same reason, his theory has very little predictive power. For a particular subject population, one can generate predictions if the problem's content taps experiences that the experimenter knows to be nearly universal or else completely unfamiliar. But for most content domains, the only prediction it can make is that responses will vary unpredictably.

The fact that the Drinking Age Problem (DAP) and Johnson-Laird et al.'s post office problem elicit high percentages of 'P & not-Q' responses presents difficulties for Pollard's theory. Most subjects have had more exposures to beer drinkers who are over 20 (legal) than under 20 (illegal) and seen more envelopes with correct postage than with incorrect postage. Thus, an implication of his differential availability view is that most subjects will choose 'P & Q' for these problems. Pollard notes this difficulty and tries to finesse it by suggesting that differential availability arises from the subject's experience of the content and context of the problem. He says:

The context relates to drinkers that are investigated by the police, or drinkers who are breaking the law, and the only available instances of these, given the context, are underage drinkers (or, in the case of the Johnson-Laird et al. study, under stamped letters). The P - not-Q link thus becomes dominant. (p. 80)

This explanation is problematic. Unless you already understand that "breaking the law" = P + not-Q, playing the role of a police
officer or postal sorter seeking violations of the law will not, in and of itself, limit your search to instances of not-Q (underaged beverage drinkers, understamped letters). This criticism is underlined by results on the post office problem for Golding's younger subjects and Griggs & Cox's American subjects. These subjects did not understand that "violating the rule" = sealed envelope + less than 20 cents postage. Playing the role of a postal sorter looking for violations did not help them one wit, even though this is the same context successfully used by Johnson-Laird et al. To look for a violation you have to know what counts as a violation. And if you already know what counts as a violation, then why not answer the selection task accordingly? Why would the relative availability of compliance versus violation episodes cause you to change your answer?

One could reframe Pollard's view of context thus: Most subjects have had experience with the police and have noted that they only question people under 20, and this makes not-Q more available than Q. But is this true? Police do not investigate guilty people only -- they query a range of people in search of the guilty. In my experience, bouncers (I have never witnessed police making such inquiries) ask to see the IDs of people who look young -- but most of these prove to be over the legal drinking age. I suspect my experience is not atypical. So all people sharing my experience of bouncers/police should choose 'Q' rather than 'not-Q'. And how many people have had any actual experience with postal sorters, to see what sorts of envelopes they pay special attention to? The point is, subjects' experience with the behavior of police and postal sorters is
bound to be as idiosyncratic as their experience with going to Boston via cab or subway. Therefore, if we reframe Pollard's view of context in this manner, responses to the DAP and post office problems should be variable, like those to the transportation problem. They should not elicit such uniformly high levels of falsification.

Last, Pollard seems to pick and choose that which he wishes to count as "actual experience." The subject, who has never been, and perhaps never even met, a postal sorter, can project himself into this role such that this imagined person's long term memory is cued. Yet this same subject cannot make the intuitive leap from haddock with water to haddock with gin. I can see no principled way of maintaining that the transportation problem and post office problem cue familiar experiences, but that the food problem does not.

Memory-cueing/ Reasoning by analogy

Memory cueing (Manktelow & Evans, 1979; Griggs & Cox, 1982; Cox & Griggs, 1982; Griggs, 1983) is a variety of availability theory that does not depend on differences in the relative availability of P & Q versus P & not-Q. Although it was first suggested by Manktelow & Evans (1979) to explain why the thematic content effect is so elusive, its primary proponents are Richard Griggs and James Cox (Griggs & Cox, 1982; Cox & Griggs, 1982; Griggs, 1983). It is a "family 4-b" explanation.

According to these researchers, subjects will falsify on the Wason selection task if they can recall past experience with:
1. the content of the problem;  
2. the relationship (rule) expressed; and  
3. a counter-example to the rule.

Recalling past experience with all three aspects of the problem allows the correct response to be "cued" from long term memory.

Unlike Pollard's differential availability theory, which requires that available disconfirming instances outnumber available confirming instances, memory-cueing theorists only require that one counter-example become available. Subjects do not* generate falsifying instances by a deductive process, but if a counter-example happens to be generated by some other means, they can recognize it as violating the rule. This highlights an important conceptual difference between differential availability theory and memory-cueing. Differential availability is an entirely nonlogical theory, whereas memory-cueing requires minimal logical competence: the ability to recognize contradiction, the most fundamental logical property.

The experiments reported by Griggs & Cox, 1982, was very important in establishing memory-cueing as a theory. The transportation and post office problems failed to elicit more logical responses from their American subjects than abstract problems did. However, 72% and 74% of subjects from the same population produced falsifying responses in two different replications of the DAP. Griggs & Cox substantiated their claim that members of their subject pool had past experience with the above three aspects of the DAP, but not with the post office problem. Thirty-three additional subjects from the same population completed a questionnaire designed to tap their

* and cannot, without explicit academic training in formal logic.
familiarity with these two rules and counter-examples to them.

The questionnaire asked:

1. whether regulations exist concerning beer and being of a certain age, and sealing an envelope and having a certain amount of postage on it; if so, then write the regulation,
2. whether they themselves had ever violated the regulation,
3. whether they could remember specific instances of someone other than themselves violating the regulation.

Only 12% wrote a rule consistent with the post office problem, but 88% wrote a rule consistent with the DAP. Only one subject recalled having personally violated the postal rule (interesting, as no such rule exists in the U.S.). In contrast, 76% of subjects reported having personally violated the drinking age rule, and 97% could recall specific instances of someone else violating it.

Griggs & Cox take this correlation of personal experience in their subject population with success on the selection task as evidence for memory-cueing theory. They also cite Golding (1981), in which older subjects who were familiar with Britain's pre-1968 post office rule did well on the post office problem, whereas younger subjects did not.

They explain the inconsistency of the results for other thematic problems (food, schools, transportation) as caused by the variable, idiosyncratic, experience of subjects with these contents. They suggest, for example, that Wason & Shapiro's (1971) transportation problem elicited higher levels of logical falsification than those of Manktelow & Evans (1979) and Pollard (1981), because Wason & Shapiro's subjects from University College London live closer to the cities named in the selection task than do the other researchers' Plymouth Polytechnic.
subjects. Therefore, Wason & Shapiro's subjects were more likely to have made a trip that happened to be a counter-example to the rule.*

Note that the fact that Griggs & Cox hazard this explanation for the transportation problem means that they only require that the subject have experience with the relation expressed by the rule. Subjects needn't have experienced the rule qua rule -- that is, as an explicit, linguistically expressed set of propositions, such as the DAP and the British postal office rule.

If memory-cueing is the full story, one wonders why performance on food problems is so uniformly low. Although memory-cueing requires that the subject have had experience with a counter-example, it does not require that the subject have had experience with the exact counter-example suggested by the uncovered not-Q card. On the DAP, for example, the subject can still be expected to choose a not-Q card that says "16 years old" even if her specific experience was of an 18 year old illegally drinking beer. The food problem studies do not report what food and drink pairs they actually used, but some authors (e.g. Pollard, 1981) have made of the fact that Manktelow & Evans' instructions used some rather odd combinations, such as, "If I eat haddock, then I drink gin". But the odder the combination, the higher the probability that a subject would have experienced

* This explanation would have difficulty accounting for Bracewell & Hidi, 1974: Even though both transportation problems were given to the same subject population, one linguistic format elicited a content effect, but the other did not. However, Griggs (1983) considers Bracewell & Hidi's instructions regarding non-reversibility too serious a confound to merit an explanation of this inconsistency.
a counter-example -- it may be true that not many people drink gin with their haddock, but I'll wager a great many have washed it down with water. The average 20 year old subject who eats three meals a day will have experienced almost 22,000 eating plus drinking events. Whatever the rules actually were, one would expect that 22,000 separate experiences would be sufficient to trigger a good number of counter-examples -- especially if many of the rules expressed odd combinations. Shouldn't the memory-cueing theorist expect a relatively consistent content effect for the food problem?

How does memory-cueing theory handle D'Andrade's Sears problem? As Griggs (1983) notes, chances are that very few subjects have been assistants to Sears' managers, or even worked in a store that required managers to authorize receipts. To handle such cases, Griggs and Cox couple "reasoning by analogy" with memory-cueing theory. Griggs (1983) points out that Johnson-Laird et al.'s British subjects did just as well on the post office problem when the stamps were Italian rather than British. He argues that this is because the familiar rule using pence is analogous to the unfamiliar rule using lire. He explains D'Andrade by saying that most subjects have probably had experience with analogous situations, such as store managers authorizing the subject's own check. "What seems to be essential is that the problem cue the subjects to recall their experience with the specific situation or analogous situations" (Griggs, 1983, p.26).

Cox & Griggs (1982) argue that they have found further support for reasoning by analogy in some experiments on transfer.
They created an "Apparel Color Problem" (ACP) which is identical to the DAP, except that the rule for the "police officer" to enforce is: "If a person is wearing blue, then the person must be over 19." Obviously, no subject has ever experienced such a rule. They gave each subject three problems to solve: an abstract problem (AP), the ACP, and the DAP. Cox & Griggs demonstrated that significantly more subjects solve the ACP when it comes after the DAP than when it comes before the DAP (75% v. 25%). Their explanation was that when the ACP followed the DAP, subjects reasoned by analogy to the DAP.

Interestingly, the ACP elicited a small but significant thematic content effect even when it preceded the DAP (ACP: 25%, AP: 4%). Griggs (1983) asserts that although the ACP does not relate directly to subjects' experience, they would have been in many natural situations in which their age constrained what they could do: drinking alcohol, driving, voting. Thus, the ACP could have cued one of these rules for some of the subjects, who could then "reason by analogy."

Unfortunately, grafting reasoning by analogy onto memory-cueing theory transforms it from a moderately specified theory into an unspecified theory. What dimensions of a situation are psychologically real for subjects? Which are most important in judging similarity? How many characteristics must be shared before a subject decides that two problems or situations are

* Cox & Griggs present other data which they also interpret as instances of reasoning by analogy, using permutations of the DAP, like "If a person is over 19 then he must be drinking beer" and "If a person is under 19 then he must be drinking coke". However, these experiments are so fraught with demand characteristics of the kind described for deformed social contracts in Chapter 2 that they are difficult to interpret.
"analogous"? These are key questions, yet they are never addressed by advocates of reasoning by analogy. Without answers to questions like these, memory-cueing/reasoning by analogy explanations are ad-hoc. In the absence of a theory of analogy, reasoning by analogy guts memory-cueing theory of its predictive value.

This can be seen by considering some possible theories of analogy. For example, are the DAP and ACP analogous because they share the same consequent term? Apparently this is not a necessary condition, because Johnson-Laird et al.'s post office problems used different terms: 50 lire stamps versus 5 pence stamps.

But perhaps problems are analogous when their consequents belong to the same class of entities* -- after all, 50 lire stamps and 5 pence stamps are still stamps. If this is the case, then why is performance so poor on food problems? There are natural situations involving explicit food rules ("If I eat red meat, then I drink red wine"; "If I eat fish, then I drink white wine"), and many involving implicit rules and relations ("If I eat breakfast cereal, then I drink orange juice", "If I eat hot chili peppers, then I drink water", "If I eat caviar, then I drink champagne", "If I eat Chinese food, then I drink tea"). These rules differ from the ones subjects were given only in the particular foods and drinks mentioned, just as the postal rules differed only in the particular types of stamps.

* Of course this begs the question. One still would need to know what dimensions are salient for deciding whether two entities belong to the same category. This formulation merely pushes the problem back one step.
The memory-cueing theorist cannot explain this difference away by pointing out that the *British* postal rule was explicitly mentioned in the task, for two reasons. First, this was not always true -- some subjects encountered the lire rule before the pence rule, and did very well, nonetheless. Second, Griggs (1983) attributes success on the Sears problem to "memory-cueing of *general* experience" (p. 25). If such general experience can be cued for check authorization, then surely it can be cued for the food problem. The same goes for the transportation problem. Isn't it likely that most subjects have favorite -- even exclusive -- ways of traveling to certain places? They walk to classes, they fly home at Christmas, they see their parents drive to work every day. Why can't they use these commonplace experiences to "reason by analogy" on the transportation problem? As mentioned above, Griggs & Cox require experience only with the relation, not with an *explicit*, rule. Unfortunately, Griggs and Cox never confront these questions.

It is quite possible, even likely, that people reason by analogy. It is even possible that this technique is only effective when combined with memory-cueing. My point is, until psychologists start developing theories of analogy, this variant of memory-cueing theory lacks any empirical content.

**Mental Models**

The mental models approach was developed by Philip Johnson-Laird (Johnson-Laird, 1982; Johnson-Laird, 1983). Explaining content effects on the Wason selection task was not his primary motivation in developing this theory. Insofar as it does account
for content effects, it relies on a form of availability. I include it because it represents a quite different view of how humans reason than do the theories previously described. Mental models is a "family 4-a" explanation.

According to Johnson-Laird, the human mind has no computational procedures that correspond to rules of inference (like modus ponens or modus tollens). Instead,
1. reasoners interpret premises by constructing an integrated mental model of them in working memory, and
2. reasoners have one piece of semantic information: A conclusion is true if the premises are true and there is no way of interpreting them so as to render it false.

These two factors can lead to logical reasoning. For example, given the premises, "Some of the scientists are parents" and "All the parents are drivers", the subject will first construct a mental model of the relation expressed by the first premise, perhaps like this:

```
scientist
scientist = parent
scientist = parent
parent
```

The first person is a scientist who is not a parent, the second and third are scientists who are parents, the fourth is a parent who is not a scientist. All four possibilities are consistent with the premise "Some of the scientists are parents." Next, the subject will try to integrate the information in the second premise into the model of the first premise:

```
scientist
scientist = parent = driver
scientist = parent = driver
parent = driver
```

This integrated mental model is consistent with two tentative conclusions: "Some of the scientists are drivers" (a valid
inference) and "Some of the scientists are not drivers" (an invalid inference). But which one will the subject choose? This is where the second factor enters the picture. According to Johnson-Laird, people know that a conclusion is true when the premises are true and there is no way of interpreting the premises so as to render it false. Therefore, the subject will search for alternative mental models that are also consistent with the premises, to see if any violate a tentative conclusion they have drawn. For example, the following two models are also consistent with the premises:

\[
\begin{align*}
\text{scientist} &= \text{driver} & \text{scientist} &= \text{parent} = \text{driver} \\
\text{scientist} &= \text{parent} = \text{driver} & \text{scientist} &= \text{parent} = \text{driver} \\
\text{scientist} &= \text{parent} = \text{driver} & \text{scientist} &= \text{parent} = \text{driver} \\
\text{parent} &= \text{driver} & \text{parent} &= \text{driver}
\end{align*}
\]

However, both render false the conclusion "Some of the scientists are not drivers." In contrast, both models are consistent with the conclusion "Some of the scientists are drivers."

Thus, mental modeling theory is very different from memory-cueing theory. According to memory cueing theory, people can recognize a counter-example as such if they happen to recall one, but they do not actively search for counter-examples. Also, in memory cueing theory people do not model the premises -- the premises function primarily as recall cues.

Johnson-Laird (1983) integrates content effects into his theory thus:

If subjects already possess a mental model of the relation expressed in the general rule, or a model that can be readily related to the rule, they are much more likely to have an insight into the task. (p. 33)

He believes that memory is important in that "no effect of content can be explained without appeal to previous experience."
Previous experience gives one a library of mental models. Realistic content makes mental models available, not mere associations.

Johnson-Laird makes no attempt to predict what kinds of content will enhance performance beyond saying that familiarity with the rule helps. However, the subject need not have experienced an explicit rule (like the DAP); he cites Wason & Shapiro's original transportation problem, D'Andrade's Sears problem, and his own fire version of the post office problem as examples. However, he provides no explanations regarding why the food problem never enhances performance, why results with the transportation and school problems are so spotty, or why results with what I have called "social contract problems" are so consistent.

Frames and Schemas

At present, explanations of content effects on the Wason selection task in terms of frames or schemas are promising, but meta-theoretical. Wason & Shapiro (1971), Wason (1983), and Rumelhart & Norman (1981) have argued that reasoning on the Wason selection task is guided by frames or schemas -- domain specific inference procedures and/or mental models. These develop content area by content area, according to the subject's personal experience. The more experience a person has had with a given content area, the more likely it is that she has acquired a frame governing inference in that area. The presumption seems to be that the processes underlying the acquisition of frames are domain general, making this a "family 1-a" explanation. However,
this view would not be compromised if most frames were built by
domain specific algorithms.

Although this view is akin to Johnson-Laird's mental
modeling theory, it is more inclusive. Schemas or frames can
enhance performance by virtue of their ability to unite the terms
of the selection task into one, unified mental representation
that can be easily manipulated via the frame's procedures (Wason
& Green (1984) present some evidence for this view using a very
simple "reduced array selection task", or RAST).* Alternatively,
performance can be enhanced via the domain specific inference
procedures that the schemas or frames embody.

The inference procedures that develop in a given content
domain need not be logical in character. In Johnson-Laird's
theory, the subject's knowledge that counter-examples are
relevant to the logical validity of a conclusion is an important
factor in rejecting tentative conclusions. In frame theory, the
subject could be judging the soundness of a conclusion using
"pieces of semantic information" that have nothing to do with
logical validity. For example, the subject's knowledge of the
social factors governing commercial transactions might guide her
response to D'Andrade's Sears problem. This knowledge can be
either declarative or procedural. Because each content area may

* The RAST is a selection task which uses only Q and not-Q cards,
and usually many instances of each. Given the rule "All
triangles are white", the subject is to determine whether it is
true by asking to inspect -- one at a time -- the minimum number
of black shapes or white shapes. The best answer is to choose
all and only the black shapes; however, one can test varying
degrees of insight into different rules by seeing if subjects
choose more confirming white shapes for one rule than for
another. The RAST is different enough from the full selection
task that results on it are not directly comparable.
have different rules of inference associated with it, a frame need not lead to a logically correct answer.

The frame theorists have not yet addressed questions like: Are some content areas more likely than others to have frames associated with them? How many experiences with a domain must one have to develop a frame? Must those experiences be of a particular kind? How does the mind parse the world into separate domains for the purpose of building frames? To what extent do different individuals share the same frames?

Without answers to questions like these, the frames explanation cannot be evaluated by appeals to empirical evidence. In principle, any content effect or non-effect is compatible with it. If a particular content elicits an effect, that is post-hoc evidence for the existence of a frame for that content domain. If it does not, that is post-hoc evidence for the lack of a frame for that content domain. The Wason selection task may indeed turn out to be a useful tool for discovering what sort of frames people have, especially if performance in certain domains is consistent across subjects, but violates logical principles. However, at present the frame view does not allow one to predict in advance which content areas will enhance performance. If one presumes that frames are built by domain general cognitive processes, then, at most, frame theory predicts that performance with the same content domain will vary, reflecting the idiosyncratic experiences of the subjects tested. But before frame theory can be considered a thoroughgoing explanation of content effects on the Wason selection task, the question of how frames are built must be addressed.
Auxiliary Mechanisms

In the early 1970s, several researchers considered the possibility that people are logically competent, but that abstract terms or relations create performance limitations (a "family 3" explanation). Wason & Shapiro (1971), Bracewell & Hidi (1974), and Gilhooly & Falconer (1974), suggested that thematic terms or relations may be more easily manipulated by auxiliary mechanisms (like working memory) than abstract terms or relations. The concrete terms used in thematic problems might enhance performance because they are more memorable than abstract symbols. A thematic relation might impose a smaller "cognitive load" on working memory if its content activates knowledge that cues the subject that the relation is non-reversible: the fact that the conditional is not reversible need not be activated as a separate and additional piece of information in working memory.

Research into this view was virtually abandoned as later results called into question the very existence of a thematic content effect. The suggestion that the superior memorability of concrete terms explains the content effect can be ruled out. The food problem has never elicited a content effect, the post office problem does not when subjects are unfamiliar with the relation, the school and transportation problems usually do not produce content effects -- yet all use concrete terms.

The hypothesis that certain thematic relations reduce cognitive load is unlikely given the spotty replication record.

* It is obvious that "All horses are animals" does not imply that "All animals are horses"; it is not so obvious that "All cards with an A on one side have a 3 on the other side" does not imply that "All cards with a 3 on one side have an A on the other side."
for the transportation problem. The transportation problem is one of the only thematic relations tested that clearly does not imply its converse. "Every time I go to Boston I travel by car" is a rather ordinary claim about a habitual way of getting to a particular destination, but "Every time I travel by car I go to Boston" sounds like the car has a mind of its own. The transportation results from the late 1970s and early 1980s have cast doubt upon the claim that a thematic relation enhances logical performance at all -- a fact that must be established before entertaining hypotheses regarding how it does this.

Before anyone realized how "elusive" content effect on the Wason selection task would prove to be, two sets of researchers -- Bracewell & Hidi (1974) and Gilhooly & Falconer (1974) -- tried to assess the relative contribution of concrete terms versus concrete relations to success on the transportation problem. Their results were contradictory.

Both sets of researchers investigated four types of problem:

Abstract Terms - Abstract Relation (AT-AR): "If there is a J on one side then there is a 2 on the other side"

Abstract Terms - Concrete Relation (AT-CR): "If I go to J then I travel by 2."

Concrete Terms - Abstract Relation (CT-AR): "If Manchester is on one side then car is on the other side."

Concrete Terms - Concrete Relation (CT-CR): "If I go to Manchester then I travel by car."

As mentioned in Chapter 2, Bracewell & Hidi also tested two different linguistic formats: "Every time P, Q" and "Q every time P."

All told, Bracewell & Hidi had eight groups (two linguistic

* The CT-CR and AT-AR rules correspond to Wason & Shapiro's (1971) thematic and abstract rules. These are the groups described in Chapter 2.
formats for each of the above four groups), with 12 subjects per group. Their results are pictured in Table 3.1 below:

<table>
<thead>
<tr>
<th></th>
<th>CT-CR</th>
<th>AT-CR</th>
<th>CT-AR</th>
<th>AT-AR</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Every time P, Q:</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>Q every time P:</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td><strong>Totals:</strong></td>
<td>11</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Number of subjects who answered 'P & not-Q'; n=12 per cell.

Bracewell & Hidi found a main effect for the linguistic format factor ("Every time..." does better), a main effect for the relation factor (the concrete relation does better), but no effect for the term factor. However, a further analysis of their data throws doubt on whether a relation factor exists at all. Although Bracewell & Hidi's data are consistent with the hypothesis that there is a relation factor, two alternative hypotheses are more strongly supported by their data:

1. There is no relation factor. Performance is enhanced only for the CT-CR problem, and only in the "Every time" format (see contrasts L1 below),

2. A concrete relation is sufficient to enhance performance, but only in an "Every time" format (see L2).

Bracewell & Hidi's hypothesis that both the relation and the linguistic format factor are important is represented by the set of contrasts L3.

\[ L_1 = +7 \ -1 \ -1 \ -1 \ \ \ \ \ \ L_2 = +3 \ +3 \ -1 \ -1 \ \ \ \ \ \ L_3 = +1 \ +1 \ 0 \ 0 \ 0 \ -1 \ -1 \ -1 \ -1 \ -1 \ -1 \ 0 \ 0 \ 0 \ -1 \ -1 \ -1 \ -1 \ -1 \ -1 \]

The sum of squares for L3, Bracewell & Hidi's hypothesis, accounts for 62% of the variance due to main effects and
interactions (i.e., of SS - SS; $F = 22.79, p < .001$, total error 1,88; effect size $r = .45$). However, hypothesis L2 (a concrete relation helps only in an "Every time" format) accounts for 69% of the variance ($F = 25.35, p < .001, r = .47$) and hypothesis L1 (performance is enhanced only by a CT-CR problem in an "Every time" format) accounts for 80% of the variance ($F = 29.40, 1,88; p < .001, r = .50$).

Thus, the hypothesis that the only cell showing an enhancement in logical performance is the CT-CR "Every time" cell -- the cell that exactly duplicates Wason & Shapiro's thematic group -- accounts for 18% more of the variance to be explained than Bracewell & Hidi's hypothesis that the relation factor exercises a separate effect, independent of linguistic format.

The efficacy of a concrete relation is further called into question by Gilhooly & Falconer (1974), whose results exactly contradict Bracewell & Hidi's. Gilhooly & Falconer investigated only one linguistic format ("Every time P,Q"), but their experiment is otherwise identical to Bracewell & Hidi's. The percentage correct for Gilhooly & Falconer's four groups is shown in Table 3.2. These figures reveal a significant main effect ($p < .05$) for the term factor (concrete terms do better), but no main effect for the relation factor, and no interactions. This directly contradicts the results of Bracewell & Hidi, who found a main effect for the relation factor, but no effect for the term factor (L3).

Indeed, the limited support that Bracewell & Hidi found for a relation factor may have been nothing more than a procedural artifact. Their unusual instruction that the conditional is "not
Table 3.2 Results for Gilhooly & Falconer, 1974

<table>
<thead>
<tr>
<th>CR</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT</td>
<td>11 : 10 : 21</td>
</tr>
<tr>
<td>AT</td>
<td>6 : 3 : 9</td>
</tr>
</tbody>
</table>

17 : 13

Number of subjects who answered 'P and not-Q'. n=50 per cell.

reversible" may have simply focused subjects' attention on the relation factor (see Chapter 2).

In short, Bracewell & Hidi and Gilhooly & Falconer provide no clear evidence for the claim that a thematic relation enhances logical performance at all, thus ruling out the hypothesis that it does this by reducing "cognitive load".*

In light of the evidence presented in Chapter 1 indicating that people do not use the basic inferences of the propositional calculus, explanations in terms of performance factors do not appear very promising. The data reviewed in Chapter 2 cast a pall on such an enterprise. Any future "performance limitation" explanations will have to explain 1) why some familiar, concrete content can be pushed through "auxiliary mechanisms" better than other familiar concrete content, and 2) why the same familiar, concrete content is sometimes processed easily, and sometimes only with great difficulty.

* One could argue that because an AT-CR rule uses abstract terms, it cannot cue non-reversibility; that it is not so clear that "Every time I go to J I travel by 2" does not imply "Every time I travel by 2 I go to J". If this were so, then the relevant cells for testing the relation factor are CT-CR versus CT-AR. Again, the results would be contradictory; these cells differ for Bracewell & Hidi, but not for Gilhooly & Falconer.
Family 2 explanations

The only family of explanation that has no representative in the literature is "family 2": Humans have a logic module, but it is only activated in answering questions within the framework of a well established theory of what is true of a content domain. On this view, people may use induction to generate hypotheses in unfamiliar domains, but once they develop some inductive confidence about their hypotheses, they test them deductively.

This explanation cannot account for the content effects reviewed in Chapter 2. Assuming that familiarity is some measure of a person's understanding of a domain, the logic module should switch on for familiar domains. How, then, could this theory account for the fact that some familiar domains elicit content effects but others do not (e.g., DAP v. food), and the same familiar domain sometimes produces an effect, and sometimes not (e.g., transportation, school)?

Other formulations are possible, but I can think of none that can handle the results of Chapter 2. For example, perhaps familiarity with the elements and relations in a domain is not enough; perhaps the logic module is activated only when the domain is familiar and the subject also has personal beliefs regarding the veracity of the relation expressed by a rule.* This explanation can also be ruled out.

One implication of this view is that people should be especially adept at evaluating the validity of conclusions when

* Even if this were true, it could not explain the results of Chapter 2. For example, the most robust and replicable content effect was for social contract problems. Yet they have no truth value; they are rules to be followed.
they have personal beliefs regarding their truth value. Van
Duyne (1976), was interested in whether people reason more
logically with sentences that express necessary truths or
contingent truths. He asked 22 subjects to generate conditionals
that they thought were "always true" (necessity condition) and
"sometimes true" (contingency condition). Each subject solved
two selection tasks that had been created from rules he himself
had generated (one necessary, one contingent).

If a logic module is activated in answering questions within
a well-established theory of what is true, then 1) Van Duyne's
paradigm should produce a substantial amount of falsification (at
least over 50%), and 2) performance should be better for "always
true" conditionals than for "sometimes true" conditionals.*
Neither prediction is borne out by the data. Levels of
falsification were low: only 6 of the 22 subjects (27%) falsified
for the "necessary truth", and 8 out of 22 (36%) falsified for
the "contingent truth." These percentages are not significantly
different, and even if they were, the inequality runs counter to
prediction. In fact, if one requires that subjects not only give
the correct answer, but give it for the correct reasons (as
assessed by verbal explanations), subjects displayed far more
insight into conditionals that were "sometimes true" than ones
that were "always true".

Even more damning to this explanation is the considerable
body of literature on "belief bias" (reviewed by Pollard, 1982),
which indicates that people do not reason more logically when

* Insofar as one's theory of what is true in a domain is better
established for rules which are "always true" than for those
that are "sometimes true."
they have personal beliefs regarding the truth value of the conclusion (see section 2.3). In such cases, subjects' performance appears to be guided, in part, by a desire or tendency to confirm their personal beliefs. When the content of a conclusion agrees with a personal belief, they judge the argument valid, and when it disagrees, they judge the argument invalid. Pollard & Evans, 1981, have demonstrated this on the selection task. Using a paradigm like Van Duyne's (1976), Pollard & Evans found that subjects were much more likely to choose the not-Q card for conditionals that they thought were "usually" or "always" false, than for conditionals that they thought were "usually" or "always" true.

These were conditionals that subjects' had generated themselves. Hence, they were familiar and subjects' had opinions regarding their veracity — optimal conditions for the activation of a logic module, according to the reformulated family 2 explanation. If a logic module is activated under these conditions, we should see a substantial amount of falsification in this experiment.

Although Pollard & Evans report selection frequencies for individual cards rather than for combinations of cards, the percentage of subjects who answered 'P & not-Q' can be estimated from the percentage of Q card selections.* At most, 8.5% of subjects falsified for "true" conditionals and 21% falsified for "false" conditionals — hardly auspicious performance for an activated logic module. This result is fatal to the "familiarity plus veracity" formulation of the family 2 explanation.

* because no one who chose Q could have answered 'P & not-Q'.

102
3.3 Summary of explanations

The theories that have been proposed in the literature represent four of the five "families" of explanation listed at the beginning of this chapter:

Frame theory is a family 1-a explanation: Humans have no logic module; instead they use rules of inference that are appropriate to the domain suggested by the problem. Current formulations presume that frames are built by domain general information processing mechanisms.

Auxiliary mechanisms is a family 3 explanation: Humans have a logic module, but auxiliary mechanisms for manipulating information create performance limitations.

Mental models and memory-cueing are family 4 explanations: Humans have no logic module, just the ability to recognize contradiction when they see it. Mental models theory falls into category 4-a, as it proposes that people actively construct models of the premises in search of ones that will refute a tentative conclusion. Memory-cueing falls into category 4-b, as it proposes that a counter-example can become available only if a person has actually experienced one -- people do not actively construct mental models in search of refutation.

Differential availability is a family 5 explanation: Humans have no logic module; rather, their performance is guided by non-inferential, general purpose heuristics.

None of these theories is satisfactory. Some are too unspecified to evaluate against empirical evidence (frames, mental models). Others are better specified (differential availability, memory-cueing, auxiliary mechanisms), but cannot
account for important pieces of evidence. To try to account for this contradictory data, some of the theories add codicils that are either theoretically unsound (differential availability), have consequences that are refuted by existing data (differential availability, memory-cueing/reasoning by analogy), or must be interpreted so loosely as to render the theory completely untestable (memory-cueing/reasoning by analogy).

None of the theories explain why social contract rules are the only thematic content to consistently elicit large content effects.

* * *

The many results cited in Chapters 1-3 demonstrate that people do not have the sort of logic module necessary for Popperian-style everyday learning. The Wason selection task is particularly interesting because it is a test of our ability to test hypotheses deductively. Although some of the theories presented in Chapter 3 provide accounts of how people can test hypotheses in the absence of a logic module (mental models, memory-cueing), these theories require that the individual bring a vast store of world-knowledge to the task.

This brings us back to the central problem: How do people acquire this world-knowledge? Is this knowledge accurate? Induction is usually conceived as a process by which the world imprints existing relations on our minds -- that is, the kinds of hypotheses it can be expected to generate describe relations between existing properties or elements.

Given that there are an infinite number of ways of carving the world into properties, and therefore an infinite number of
relations between properties to serve as hypotheses, we must generate an enormous number of incorrect inductions. Yet results on the Wason selection task show that we are very bad at testing descriptive rules -- the very sort of hypotheses that induction provides. How, then, do we weed out all these incorrect inductions?

More puzzling: If the evolutionary purpose of human learning is to provide valid generalizations about the world, then surely the need to detect violations is greatest for descriptive rules. Why, then, are we so bad at detecting violations of descriptive rules, but so good at detecting violations of social contract rules? Social contract rules do not describe the way things are; they do not even describe the way existing people behave. They prescribe: They communicate the way some people want other people to behave. They are rules to be followed. One cannot assign a truth value to them. Why, then, do we appear to reason logically in response to social contract rules, but not in response to descriptive rules?

These are the questions that motivate the remaining chapters.