THE UNIVERSITY OF
WARWICK

University of Warwick institutional repository: http://go.warwick.ac.uk/wrap

A Thesis Submitted for the Degree of PhD at the University of Warwick

http://go.warwick.ac.uk/wrap/4040

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

Confirmation Bias and the Testing of Hypotheses about Other People

Brendan Joseph Burchell

A thesis submitted in
partial fulfilment of the
requirements for the degree
of Doctor of Philosophy

at the

University of Warwick

Department of Psychology

September 1986

Title. page I

Contents

List Of Tables

List of Figures

Acknowledgements

My Supervisor, Dr Ian Morley, has undoubtedly been the greatest guiding force in my academic training, and I am very grateful to him for that. I have also received support from virtually all of the other members of staff in the psychology department of Warwick University at some point or other during my studies there. In addition, I would like to thank Dr Don Pennington for encouraging me to undertake this PhD in the first place.

Lynda, my wife, has put a tremendous amount of effort into helping me to finish this thesis. In particular, I am thankful to her for reading the draft manuscripts and contributing greatly to the quality of the finished product.

The other people who deserve my gratitude are those who have encouraged me throughout the last five years and in particular in the final writing of the thesis. My parents, Charles and Bernadette, and the team I am currently working with at the University of Cambridge were particularly supportive and helpful in this respect.

September 1986

Declaration

The material presented in this thesis is entirely the work of the author with the following two exceptions:

1/ The design and data collection of part of the experiment presented in Chapter 4 was conducted by Dr D.P. Pennington.

2/ Some of the data collection for the experiment reported in chapter 5 was done by Ms. Kaye Trattles for her final year dissertation.

Some of the empirical findings and conceptual arguments presented in this thesis have also been published or presented to conferences in one or more of the three articles authored by myself and listed in the reference section.

Summary

Critical reviews of the literatures on self-fulfilling prophecies and self-confirming hypotheses uncovered several weaknesses in key works on those topics. In particular two important flaws were revealed. Hypotheses and expectations were confused and confounded and the most important aspect of these effects in person perception, changes in the perceiver's representation of the target, were ignored. Instead these works either made inferences about the perceivers' judgments from other individuals with different perspectives, or claimed to have demonstrated the effect of manipulating the hypothesis whereas their results were probably attributable to manipulating expectancies instead. It was argued that both of these types of inferences are invalid, and re-analyses of data from empirical works showed that the claims were not justified.

A series of experiments was conducted in an attempt to find unequivocal evidence of self-confirming hypotheses. Numerous reasons were found as to why the phenomenon was highly unlikely to occur in social interaction. For instance, the asking of biased questions was found not to occur when perceivers generated their own questions to ask instead of selecting from a list given to them. In addition, subjects modified the questions they asked during the course of social interactions in such a way as to eliminate any possible bias in information search. Even when questions searching for confirmatory evidence were asked there was little evidence that interviewers' judgements were biased in favour of confirming their hypotheses.

By contrast strong evidence was found for self-confirming hypotheses when subjects used information from their own memories to test hypotheses about aquaintances.

These findings were discussed in the light of other paradigms within social psychology. Reasons why social cognition has, at times, so underestimated human rationality were considered and several conclusions were made including the need for greater caution in attempting to emulate and understand social processes in a laboratory setting.

Chapter 1

Self-Fulfilling Prophecies In Person Perception

Introduction.

The phenomenon named the self-fulfilling prophecy has been the subject of a considerable amount of research in the social sciences and, in particular, in social psychology. This chapter will provide a review of the social psychological literature on self-fulfilling prophecies. Three empirical studies of different self-fulfilling prophecies will be described in depth. Using these examples to illustrate the arguments, new criteria for evaluating self-fulfilling prophecies will be proposed, paying particular attention to the empirical studies in person perception that claim to have demonstrated the self-fulfilling prophecy phenomenon. It will be argued that the essential features of a self-fulfilling prophecy are different in the study of person perception than in other areas of the social sciences. Whereas in other applications of self-fulfilling prophecies the primary interest has been in the target of the prophecy, in person perception it should be on the perceiver's representation of the target of the prophecy. The experiments purporting to show self- fulfilling prophecies in person perception have failed to take this into account.

The framework for analysing self-fulfilling prophecies in person perception developed in this

chapter will then form the basis of the analysis of the closely related phenomenon, the self-confirming hypothesis, which will be the primary focus for the theoretical and empirical work presented in the rest of this thesis.

The discussion of self-fulfilling prophecies will start, however, outside of social psychology with the work which introduced the concept of the self-fulfilling prophecy into the social sciences. Merton's classic paper called "The Self-Fulfilling Prophecy", published in 1948 (and reprinted as a chapter in his 1957 book), continues to be cited widely and includes the definition of the self-fulfilling prophecy that is still the most commonly used today.

Merton's original conceptualisation of the Self-Fulfilling Prophecy.

Robert Merton was the first writer to explicitly define and use the concept of a self-fulfilling prophecy. Although he was himself a sociologist, his most memorable and most quoted example of a self-fulfilling prophecy is taken from an actual instance of a real economic incident that occured in the year 1932.

The Last National Bank was a thriving institution, and was doing at least as well as most American banks. Then, for no apparent reason, the manager noticed that they were doing a very brisk trade one particular

Wednesday, usually a quiet day. Two dozen men from the local factory were all queuing up to withdraw their money. Instead of subsiding to a normal level, the queues grew all day as the men became more anxious to withdraw their money. The next day saw even longer queues, and the bank was eventually unable to meet the demands of its creditors -- it was suddenly bankrupt.

This was due to what Merton calls the "Thomas Theorem". Thomas (an influential sociologist in the earlier half of this century) stated that "If men define situations as real, they are real in their consequences" (1928). The investors had defined the pending closure of the bank, and it was this definition alone that caused the reality -- the closure of the bank. This shows a very stark contrast between the social world and the physical world. Prophecies, rumours or predictions could not make the slightest difference to physical realities such as the boiling point of a liquid or the appearance of Halley's comet in 1986. The exact opposite is true of phenomena in the social world -- whether at the level of the individual deciding to ignore a person they *suspect* of disliking them or at the level of the government launching an attack because of a *perceived* threat or a stockmarket crashing because of *speculations* about falling prices. These situations all have one thing in common -- the reality of the situation has changed because of the way in which an individual or a group has defined that situation or prophecied a particular future for it.

Merton defined self-fulfilling prophecies thus:-

"The self-fulfilling prophecy is, in the beginning, a <u>false</u> definition of the situation evoking a new behaviour which makes the originally false conception come <u>true</u>. .... The specious validity of the self-fulfilling prophecy perpetuates a reign of error. For the prophet will cite the actual course of events as proof that he was right from the very beginning. Such are the perversities of social logic."

(1957, p. 477, emphasis in original).

Merton goes on to explain how this effect may be behind many of the ills of society and the cause of the persistence of some processes and beliefs. For instance, the Negroes were excluded from early trade unions in North America because, it was reasoned, they are not used to the traditions of organised labour and would thus be poor union members and undercut accepted rates. Because of this discrimination the Negroes were forced to take jobs at low rates of pay, undercutting the unions. This gave the Whites hard evidence that the Negroes were, indeed, unworthy of trade unionism and that the decision to exclude them was justified all along. The white citizens did not realise that <u>it was their own actions that created these facts</u> rather than the inherent nature of the black workers.

The Self-Fulfilling Prophecy in Social Psychology.

Many studies paralleling these findings have been translated into fields that are of direct relevance to social psychology such as (for example) person perception (Kelley, 1950), self-perception (Swann & Hill, 1982), racial prejudice (Word, Zanna & Cooper, 1974), sex stereotyping (Zanna & Pack, 1975), social stereotypes (Snyder Tanke & Berscheid, 1977), attitude change (Lord, Ross & Lepper, 1979), competitive and cooperative styles of interaction (Kelly & Stahelski, 1968; Snyder & Swann, 1978b), educational attainment (Rosenthal & Jacobson, 1968, Swann & Snyder, 1980) and even experimenter effects (Rosenthal, 1976). To form the basis of the arguments that are to follow later in this chapter, three of these experiments will now be described in detail.

The Pygmalion Studies: Self-Fulfilling Prophecies in the classroom.

This very influential study, conducted in an educational setting, was reported by Rosenthal & Jacobson (1968). They set out to demonstrate experimentally the mechanisms by which teachers' expectations of their pupils' future performance (their prophecies) could fulfil those expectations. Although aspects of the methodology of this work were criticised initially, it has stood the test of replication several times over (See,for instance Crano & Mellon, 1978; Seaver, 1973 or

Sutherland & Goldschmid, 1974), and it is typical of the many programmes of research that followed it.

Rosenthal and Jacobson were investigating the reasons why children from disadvantaged ethnic minorities perform so poorly in the education system. Several obvious reasons have been put forward for this: the disadvantaged home backgrounds; language problems; motivational problems and so forth. Rosenthal and Jacobson started looking at the problem from the other point of view -- maybe it is the teachers' behaviour towards the children that cause the children to fail rather than some feature of the children themselves. It would be very difficult to manipulate teachers' expectations of pupils from ethnic minorities within the confines of a controlled experiment so instead the teachers were lead to believe that some of their pupils (in fact a random sample) were likely to show a marked improvement over the next year. Would this (bogus) expectation for some pupils effect their actual performance?

The teachers were told that their class was to help in the validation of a new test designed to predict academic "bloom" or intellecteal gain in children. The test papers were bound in impressive-looking folders, and bore the title *"Test of Inflected Aquisition"*. They were, in fact, no more than a novel type of intelligence test that the teachers were unlikely to have been familiar with. Soon after this test was

administered the teachers were told in a casual *"By the way, in case you're interested ....."* way which of their pupils were likely to "spurt" in the next 12 months. In fact, these "spurters" were five pupils picked at random from each class.

All of the children were given several follow-up I.Q. tests over the following year and a half. These tests revealed that those children in the experimental condition had shown considerably greater gains over that time than the pupils in the control groups. The teachers' expectations and nothing else had caused those gains in intelligence. It is all the more impressive that the gains were detected by objective measuring devices, not just in the eyes of the teachers.

The mechanisms for these increased gains in I.Q. are interesting. They were not simply caused by the teachers devoting more time and effort to those pupils. If this were the case then one would expect those children to have gained most on the verbal component of the test; in fact the experimental group showed a greater differential improvement on the reasoning tasks. It seems as if the communication of the expectations took place unwittingly using subtle non-verbal channels. It was not only the possibility of this effect that made this study so influential in education, but also its magnitude. The differences between the experimental and control groups were larger than had been achieved from even successful (but costly) interventions to boost

educational attainment.

To summarise, what began as deliberately bogus information from the experimenters was transformed into expectations on the part of the teachers, and it was <u>these expectations alone</u> that caused the initially false information to become true.[1]

1/ It is of interest to note that the exact opposite of a self-fulfilling prophecy can also occur -- Merton calls this a suicidal prophecy. For instance, environmental scares that prompt quick action by governments may succeed in solving the problem before it develops, leaving the government open to the charge that the action was unnecessary. A neat laboratory demonstration of a suicidal prophecy was conducted by Swann & Snyder in 1980. Instructors who were told to devote more time to teaching allegedly low ability pupils caused those pupils to outperform the allegedly high ability pupils.

Example 2: Self-Fulfilling Prophecies in Inter-Racial Job Interviews.

Word, Zanna & Cooper (1974) investigated the presentation of black and white job applicants to white interviewers. Their review of the literature comparing communication with members of stigmatised groups (for instance the physically handicapped) and with communication with other "normal" individuals. They found that people adopted more distant and less immediate nonverbal behaviours in talking to the stigmatised groups (Kleck, 1968). They also found evidence that interactants reciprocate the nonverbal styles of their partners (Rosenfeld, 1967). Putting these two findings together, Word, Zanna and Cooper proposed that the same two phenomena may occur when a white job interviewer interacts with a black applicant. They used two separate experiments to test two hypotheses. Firstly, they hypothesised that white interviewers would display less immediate nonverbal behaviour when interviewing black rather than white job applicants. Secondly, they proposed that anyone faced with this style of interaction would be induced into a less immediate style themselves, which would in turn cause them to be perceived as less desirable individuals and less suitable for employment.

The subjects in the first experiment were 14 white male undergraduates. Through a series of experimental manipulations they were led to believe that they were to

perform a task in teams of four. They were introduced to two other "subjects" (in fact confederates of the experimenter) to be in their problem-solving team. The subjects were then told that their first task was to interview four high school pupils to select one more member of the team. The instructions, plus promises of financial rewards for sucessful teams, ensured that the "interviewer-subjects" took their task seriously and made a real effort to select the applicant who was, in their own eyes, most able. They then interviewed three pupils before they were debriefed. The first interview was discounted, leaving two interviews for experimental analysis; one of these interviews was always with a white pupil and one with a black pupil.

The high school school pupils were chosen from a group of five -- three white and two black. The interviewers were given a list of 15 questions to ask. Prior to the experiment the school pupils had all been trained to answer these questions so as to appear equally competent and their nonverbal communication was also standardised as far as was possible.

The dependent variable for this experiment was the immediacy of the interviewers' nonverbal communication. This was measured in several ways such as the interviewers' eye-contact and their seating position. These indices were added together using standardised weightings to give an overall measure of immediacy.

Analysis of the data showed that the interviewers were behaving in a less immediate way with the black interviewees than with the white interviewees. In addition the interviews lasted 25% longer with the white interviewees; the interviewers also made fewer speech errors per minute -- a further sign of a less formal and more friendly approach with the whites.

Thus the first hypothesis was confirmed. The second hypothesis could then be tested. What differences would these two styles of nonverbal communication have on the behaviour of interviewees and the way that behaviour would be interpreted?

To answer this question a second experiment was conducted. Two white confederates were trained to act as interviewers, both acting in both the immediate and non-immediate conditions. These conditions were designed to be as similar as possible to the averages of the interviewers' nonverbal behaviour in the "white interviewee" and "black interviewee" conditions from the first experiment. Thirty subjects were to act as interviewees and were told to pretend they were job applicants. Again there were financial incentives to induce the subjects into taking their task seriously.

These interviews were video-recorded and played back to a panel of two rater-judges at a later date. These video tapes (<u>of just the interviewee, not the interviewers</u>) were rated on several scales to indicate

the interviewees' perceived performance and aptitude. Two other measures were taken -- a standardised aggregate of nonverbal immediacy and the interviewees' mood and their opinion of the interviewer after the interview.

All three of these categories of measures showed support for the hypothesis. The interviewees in the less immediate condition were rated by the judges to be less adequate for the job and were also rated as less calm and composed. Analysis of the interviewees' nonverbal behaviours showed that they were reciprocating the interviewers' nonverbal cues, returning less immediacy in the less immediate condition. Finally the interviewees liked the interviewers significantly less and rated them as significantly less adequate in the "non-immediate" condition, but their lower mood was not significantly different from the "immediate" condition.

Word, Zanna & Cooper argue that the complex mechanisms that operate to perpetuate the disadvantage of stigmatised racial or ethnic minorities can now be better understood. White interviewers may, even if they want to give black job applicants a fair assessment, unintentionally behave in a more formal and careful manner with them. This will, in turn, make the black job applicants perceive the interviewers as less friendly, and they will respond in a similar formal manner. This will endear them less to the interviewer, who is more likely to favour the white applicants. The

interviewer will come away satisfied that he has given all the applicants a fair chance, but the black applicants simply do not present themselves so well -- after all, he saw it for himself with his own eyes!

The authors also say that it is now possible to account for persistent discrimination in the labour market without recourse to motivational expla nations involving bigotry, ethnocentricity, ego-defence or projection; it is simply a self-fulfilling prophecy. The white interviewers' beliefs that black job applicants may be less skilled than their white peers has created a world in which this has actually happened.

Word, Zanna & Cooper go on to argue that, unlike many other errors in person perception, this one is unlikely to be self-correcting. With every interview with a white applicant who in induced into performing well, and with every interview with a black applicant who is encouraged to perform badly, the interviewers' expectations are likely to become stronger than before and even more likely to perpetuate the effect in the next interview. This undesirable phenomenon is not necessarily caused by any malicious intent on the part of the discriminator, nor necessarily by any deficiency on the part of the discriminated, but can arise through the complex process of interaction between the two parties.

Example 3. Self-Fulfilling Prophecies Confirming the Attractiveness Stereotype.

Another study to investigate the mediation of social stereotypes was conducted by Snyder, Tanke & Berscheid in 1977. They were looking at the very widespread stereotype that *"What is beautiful is good"*. Both empirical findings (eg. Dion, 1972 and Dion, Berscheid & Walster, 1972) as well as folklore (eg. Cinderella and the ugly sisters) suggest that more attractive individuals are perceived to be better people on a diverse variety of traits from romantic love to intelligence. It seems clear to psychologists that even if there were to be a kernel of truth in any of these beliefs (and empirical evidence by Goldman & Lewis in 1977 supports the view that there is), in reality the difference between attractive and unattractive individuals is much smaller than the stereotype suggests. Why then does the stereotype persist so consistently over the life-time of the individual and between generations? Why don't we ever learn?

Snyder et al proposed that the answer to this question lies in the fundamental difference between physical knowledge and socially aquired knowledge. Most of the current research into person perception focuses on the way individuals process information that is given to them by the experimenter. This is particularly true of attribution theory research. Snyder, Tanke & Berscheid point out, however, that this information must

be obtained by the individual before processing can take place. Furthermore the gathering of information can be as difficult a task and thus as likely to cause error and bias as the processing of that information.

Snyder, Tanke & Berscheid's experiment set out explicitly to look at how styles of interaction may effect the types of information we elicit from others, in turn effecting the attributions we make about them. If, for instance, we are pleasant, warm and animated towards someone, they will probably reciprocate in a positive manner, and we may well conclude that they were genuinely a nice person. Conversely, if we held negative expectations about others, we may act negatively towards them, to which they will respond negatively which will in turn be taken as evidence that the initial negative expectations were indeed right all along. Snyder et al (1977) used the attractive/ unattractive dimension to explore the way in which our perception of others can lead eventually to them behaving in a manner consistent with that perception, even if our initial perception was entirely false.

Two subjects took part in each interaction in the experiment, a male "perceiver" and a female "target". They were told that they were to take part in an experiment that looked at the role of nonverbal communications when people get to know each other, and they had been selected to take part in the telephone type of conversation. They they were also told to come

to different rooms on separate corridors for the experiment. This ensured that the two conversants did not see each other at any point before or during the experiment, which was essential for the manipulation of perceived attractiveness.

The two participants were then told that to help the conversation flow more naturally they would be given the opportunity to find out a bit about each other. Both the male and the female were then given a biographical inventory to complete by giving details of their education, hobbies and so forth. The males were then told that it helped if one knew what the other person looked like, because this would allow one to picture them more easily. A Polaroid instant photograph was then taken of them. When the males received the female's biographical inventory, it also had a Polaroid photograph attatched to it.

This was not, however, a photograph of the female subject. It was one of a number of photographs of a female that had been prepared previously depicting either a woman that was considered to be very attractive or of a woman that was considered to be very unattractive by a panel of rater-judges. The female subjects in this experiment were not actually photographed, and no mention was made to them of photographs.

In order to check that this manipulation had produced the desired effect on the males, both subjects

were asked to rate their partners on a total of 34 bipolar adjective scales, using the biographical details. As expected the addition of the attractive rather than the unattractive photograph made the males rate the females more positively on a number of scales.

To monitor the interactions in detail a stereo tape recorder was used so that the male and female subjects could be recorded on separate channels. The female channels of the recordings were then rated separately by a group of 12 rater-judges, naive as to the hypothesis being tested in the experiment or the condition of any particular dyad. Similarly, a group of nine rater-judges listened to just the male track of the tapes. The rater-judges scored each interactant on a total of 44 bipolar scales, which included the 34 scales that the conversants used to rate each other.

The analysis of the results showed quite clearly that the allegedly attractive females were judged to be reliably more socially desirable than the females in the "unattractive photograph" condition on a wide variety of scales.

Beyond this, a more detailed analysis of the results becomes difficult because of the large number of dependent variables (in excess of 50 at times). However, from a post-hoc test Snyder et al conclude that the attractive females were only more likely to show more positive behaviours on those scales where the

stereotyped views of the males had predicted a difference. The calculations that lead to these conclusions were as follows. Earlier in the data analysis a discriminant analysis had been employed on the 34 adjective bipolar scales the rater-judges used to rate the females' behaviour. Twenty-one of these scales were identified as high discriminators by virtue of having a difference of 1.4 or more standard deviations between the means on that scale for the attractive verses the unattractive conditions, the other 13 were rejected as low or non-discriminators (see, for example, Tatsuoka, 1971). Looking back to the males' intial stereotyped impressions of the females from the bigraphical details and the photographs, 17 of the 21 high discriminators were in the predicted direction, but only eight out of the 13 low discriminators were in the "attractive is positive" direction. Since binomial tests show the first proportion as highly unlikely to have occured by chance ($p=0.003$), and the second as very likely to have occured by chance ($p=0.29$), Snyder et al claim that the males only found what they were looking for; what occured was not an all-embracing halo effect. In other words, the males were likely to make the allegedly attractive females behave in a more sociable way because they expected them to be more sociable, but because the males did not expect them to be more intelligent or sensitive, they were not seen as more intelligent or sensitive by the rater-judges.

The next analysis to be performed in unravelling the complex process attempted to find out how the males mediated the behavioural confirmation. Again, the fact that there were more dependent variables (50) than the number of subjects (male perceivers, n=38), made a straight forward multivariate analysis of variance inappropriate so Snyder et al settled for fifty separate univariate analyses of variance. Twenty-one of these proved significant at or above the 0.05 level. From an inspection of those 21 scales it was concluded that the males were generally more positive in their interactive styles when they were lead to believe that their female partners were more attractive.

After the conversation the female partners were asked how they thought the males had perceived them. In the attractive photograph condition, the females reported that the males had treated them in a manner more typical of the way in which they were normally treated. The females also believed that the males had perceived them more accurately when their male interaction partners had been under the impression that the woman was physically attractive.

Snyder, Tanke & Berscheid claim to have identified another case in which a perception of a situation has left its mark on the situation. They claim that the males' erroneous expectations of the females have made the females act exactly in accordance with those expectations.

A follow-up experiment goes further in exploring this expectancy-confirming phenomenon. In case it might be claimed that Snyder, Tanke & Berscheid simply demonstrated the strength of sexist beliefs of males, Andersen & Bem (1981) enlarged the paradigm to include female perceivers talking with allegedly attractive and unattractive males, and both male and female same-sex pairs. The effect was just as strong in all conditions. The only individuals for which the effect was not replicated in an identical way were the ones classified as androgynous, individuals who do not encode and organise information so readily in terms of sex-linked associations. Androgynous females were found to be unlikely to find confirmation of the "what is beautiful is good" stereotype, but apart from that the effect proved itself equally powerful across all conditions.

In their discussion of these findings, Snyder et al argue that the same type of effect may account for several phenomena of both theoretical and practical importance. For instance, one disturbing finding from the psychology of the physically handicapped is that even very able individuals soon fit in with the cultural stereotype that expects them to be highly dependent. Snyder, Tanke & Berscheid point out that their results suggest that this may be caused by people acting towards the disabled as if they were dependent (in the same way as the males were acting as if the "attractive" females were more socially warm), and the handicapped

individuals would eventually find themselves behaving in a manner consistent with those expectations.

Another example will show the potential theoretical contribution of this experiment. One of the most pervasive findings in attribution theory research is that individuals attribute the behaviour of others to their dispositions very readily, at the expense of making situational or external attributions. So pervasive is this attributional bias that it has been named the "Fundamental Attribution Error" (Ross, 1977). All previous explanations of this error[2] leave one question unanswered; why didn't people learn to correct this error after being proved continually wrong by the lack of cross-situational consistency in behaviour?

If, however, the people we interact with really do come to behave in accordance with our expectations, then it is not surprising that we see them as being so consistent. What we would not realise is that this consistency is caused by the way we act towards them, and they may appear very differently to others who have different expectations of them. We have very little chance to find this out, however, since we rarely have the opportunity to observe individuals we know interacting with others. Even when we hear other people reporting that they think a particular other is, say, very friendly when we have found them to be unfriendly, we can dismiss this by saying that others are not such a good judge of character as ourselves!

It should be apparent from these three examples of social psychological self-fulfilling prophecies that the phenomenon is not just an interesting curiosity or an eloquent example of what can happen in social interaction. It is a phenomenon that would seem to account for much of the error in social perception in everyday life and, most of all, for the perpetuation of erroneous beliefs once formed.

An Attempt at a taxonomy of Self-Fulfilling Prophecies.

Darley and Fazio published a review article in 1980 in an attempt to unite and make sense of the growing volume of reaearch into self-fulfilling prophecies in social psychology. A simple model of social interaction was proposed in order to catalogue expectancy confirmation processes and highlight gaps in current knowledge.

2/ For instance, the saliance of individuals compared to situations, or the greater representativeness of of people as opposed to situations as causes of behaviour.

Darley and Fazio's model of social interaction identifies six crucial stages that are useful in understanding how self-fulfilling propheces may occur. They immediately point out how the model has two main weaknesses. First, most social interactions do not have discrete beginnings and ends; the information and impressions from one interaction shape the next interaction, so it is maybe better seen as a cyclic rather than a linear process. Second, the labelling of one participant as the perceiver and the other as target is oversimplistic since more normally both participants are simultaneously attempting to form a better impression of the other while monitoring the impression the other is forming of them. These criticisms aside, the six stages of their model of social interaction are as follows:-

A/ The perceiver forms an expectancy about the target

B/ The perceiver acts in a manner congruent with this expectancy

C/ The target interprets this behaviour

D/ The target responds

E/ The perceiver interprets the target's response

and finally,

F/ The target interprets his or her own response.

Darley and Fazio (1980) use this model as a framework to describe each of these phases in detail and demonstrate how biases may occur at each of these steps.

This chronicling of the stages does not, in itself, shed much light on different types of self-fulfilling prophecy. It will be argued here that a more detailed understanding of self-fulfilling prophecies is needed, and that this will reveal that some of the claimed demonstrations of self-fulfilling prophecies fall short of a more useful definition of the phenomenon.

The one distinction that Darley and Fazio do make is between, (a) situations in which the behaviour of the target person is not actually altered by the perceiver but is misinterpreted by the perceiver as confirming the expectation and, (b), situations in which the perceivers' expectations have a direct bearing on the target and actually alter the target's behaviour to bring about the expected behaviours. An example of one of each of these two types of self-fulfilling prophecy will clarify this distinction.

A good example of individuals perceiving what they expect to perceive was demonstrated by Kelley in 1950. Students were given a description of a visiting lecturer, for half of the students the description contained the word warm, the other half of the students received the same description but with the word cold substituted. After attending the visiting lecturer's class, the subjects in the "warm" condition actually came to rate the lecturer as being warmer and also as being more positive on a variety of measures. It seemed

as if the two classes had interpreted the same actions as being evidence of different dispositions.

Snyder, Tanke and Berscheid (1977) went to great lengths in an attempt to prove that this bias in interpretation did not account for the confirmation of the expectancy in their experiments. The naive rater-judges were employed to demonstrate that the females' behaviour was different in the attractive and unattractive conditions, in a way that would confirm the initial expectation. Thus they claimed to have demonstrated an actual difference in the targets' behaviour caused by the perceivers' actions.

Apart from Darley & Fazio's paper, the only other recent work that attempts to bring together and integrate the literature on self-fulfilling prophecies was Jones' 1977 book entitled *"Self-Fulfilling Prophecies: Social, Psychological and Physiological effects of Expectancies"*. The book draws upon a very wide range of applications of self-fulfilling proph-ecies, such as medical evidence, placebo effects, labeling theory, academic performance and achievements of goals. The book attempts to integrate much of the diverse anecdotal and empirical evidence under a unified model of goal setting and allocation of effort in the pursuit of that goal. This analysis can explain and lead to a greater understanding of some of the self-fulfilling prophecies reported, but contributes little to most social-psychological applications. In fact,

apart from general discussion of the social psychological literature on topics such as implicit personality theories and the origins of expectancies, no real attempt is made to argue whether and under what circumstances self-fulfilling prophecies will occur. It is not even apparent that there was any benefit in integrating such a diverse mixture of material in one book if a framework cannot be found to link it.

Change in the Target or the Perceiver's Representation of the Target?

The first distinction that will be made between self-fulfilling prophecies concerns the nature of the matter that has been confirmed. In the "Pygmalion in the classroom" experiment the expectation of the teacher (the perceiver) caused a change in the aptitude of the pupils (the targets) as measured outside of the direct influence of the teachers. This change is thus both objectively observable and enduring. Contrastingly, the changes brought about in the targets in the other two examples discussed here were presumed to last only as long as the interaction. While the targets were observed to behave differently depending upon the expectations of the perceivers, the focus of interest was the way in which the perceivers and other observers interpreted that behaviour. These two types of self-fulfilling prophecy can thus be called categorised into what will be called "enduring change in target" and "temporary change in target" types. While enduring

changes in the target may seem to be of greater importance, they are probably less likely to occur in the brief encounters observable in laboratory experiments. The impressions other people form of the target are also very important to their destiny though. Take, for example, the plight of the black job applicants emulated in Word, Zanna & Cooper's experiment. The failure to be offered employment could have serious long-term effects for them.

It is, of course, possible for an interaction to have both an enduring effect on the target and on the perceiver's impression of the target. Swann & Hill (1982) demonstrated that even a fairly brief laboratory interview could effect a target's self image as measured several days later. Nevertheless, all of the experiments that have investigated self-fulfilling prophecies can be classified according to whether the primary focus of interest was an enduring change in some characteristic of the target, or a change in another person's representation of the target. The latter of these two types is the primary interest of self-fulfilling prophecies in the study of person perception.

Perceivers and Naive Observers

This type of self-fulfilling prophecy in which the primary interest is on the inferences made from the target's behaviour while influenced by the perceiver's expectations can be further sub-divided depending on who

is experiencing the self-fulfilling prophecy. The two person perception experiments that were described both measured not the perceiver's representation of the target, but the representations of the "observers" who rated the target after hearing or seeing the target during the interaction with the perceiver.

The main reason for employing these observers was presumably to illuminate the actual mechanism of the self-fulfilling prophecy. More specifically, the primary contribution of the observers (or rater-judges) was to be able to differentiate between Darley & Fazio's two types of self-fulfilling prophecies, the "actual confirmatory behaviour" type verses the "perceived confirmatory behaviour" type discussed earlier. Beyond an interest in the technicalities of self-fulfilling prophecies, though, the perceptions of the interactor are of far greater practical importance as far as self-fulfilling prophecies impinge on our everyday lives. If we need to get to know a person it is much more likely that we would talk to them ourselves rather than passively observing a third party conversing with them. There are clear exceptions to this; for instance a large part of our impressions of television personalities and politicians comes from hearing them being interviewed by others.

It is worth noting in passing that in both of the experiments discussed here the observers did not share the perceivers' representation of the target. It is

also much more likely that, in a given real-life situation that two individuals would share, at least to some extent, the same representation of the target. This would, of course, have an important but unexplored effect on the representation that the observer gains of the target. As Moscovici (1981) has pointed out, psychologists have often ignored this great consensus between individuals from the same cultural background in their attitudes and stereotypes.

In the next section it will be argued that the position of the perceiver who holds the expectancy and the observer, naive to the expectancy, are completely different. Because the information available to the observers in the two experiments was controlled in an artificial way, the experiments actually tell us little or nothing about the "observer experienced" self-fulfilling prophecies. Furthermore, because of a failure to consider fully the information available to the perceivers in these interactions, it will also be argued that what has actually been experienced by the perceiver falls short of any useful definition of self-fulfilling prophecies in person perception.

Conceptual and Methodological Criticisms of the Self-fulfilling prophecy experiments.

A. The "Observer experienced" Self-fulfilling prophecy

The two demonstrations of self-fulfilling prophecies in person perception arising from social interactions described above (Word, Zanna & Cooper, 1974 and Snyder, Tanke & Berscheid, 1977) both used rater-judges to observe the targets in the interactions. The video cameras used by Word et al were positioned to show just the interviewee and not the confederate-interviewer. Similarly Snyder et al allowed the rater-judges to listen only to the females' voices in the conversation, not the voices of the males who were eliciting their more or less sociable behaviour. The logic in this was, presumably, to ensure that the expectancy was not mediated directly from perceiver to rater-judge, but only through the behaviour elicited from the target. Far from providing a stronger proof of the power of the expectation on the target's behaviour, this manipulation has instead made the observer's role so artificial as to take away any usefulness of their role in the experiments.

Consider the following (albeit symplistic) example as an illustration of a situation in which a perceiver may generate evidence from a target that would appear in one way to a naive observer who could not observe the

perceiver's side of a conversation but in a very different way from the point of view of the perceiver.

A female student (the perceiver) wants to impress her male lecturer (the target). She does this by paying him many compliments and adopting an immediate nonverbal manner. From the discussion so far it would come as no surprise if the lecturer reciprocated by also adopting a friendly manner. To a naive observer observing only the lecturer this would be taken as evidence of a friendly disposition on his part. However, an observer looking at both the interactants, or the student herself, may arrive at quite different conclusions from the lecturer's behaviour -- they could both marvel at her powers of influence, and be fully aware that most people who were complimented would adopt a pleasant manner (there is, in fact, well documented evidence (Singer, 1964) that more attractive females and Machivellian males do knowingly manipulate others using exactly these mechanisms).

To put this into the terms used by attribution theorists, the student has recognised that it was the situational forces acting on the lecturer and not his personal dispositions that were responsible for his behaviour. On the other hand, an observer who had not observed the situational forces acting on the lecturer would have little option but to make a dispositional attribution and put the lecturer's freindliness down to his good nature. It is logically possible that

Snyder et al's male perceivers made the same attributions as the imaginary student -- "*Of course she sounds interested and sociable, I was being as friendly as I could to chat her up*", or "*She sounded a bit bored, but that was probably because I wasn't bothered myself*".

The words <u>logically possible</u> are used because previous attribution research literature has shown that individuals are very prone to ignoring even blatant situational forces and still make dispositional attributions (eg. Ross, Amabile and Steinmetz, 1977). It is clear, however, that any attribution that is made after observing only one side of a conversation is highly likely to produce an even more dispositionally biased set of attributions. It cannot be readily inferred from this, though, that those attributions would have been the same if the observer had been able to observe both sides of a conversation. Since it is such a highly unusual situation to observe just one side of a conversation (say when one overhears a person on the telephone) it is a phenomenon of little interest. And clearly from the senario given above, it is impossible to make inferences from one situation to the other.

Would the rater-judges still have thought that the females in the attractive photograph condition were more likable if they knew that the males were being so much nicer to them? Would the rater-judges still have rated

the interviewees in Word,Zanna & Cooper's "*non-immediate*" condition as less suitable for the job if they knew that the interviewers had given them such a hard time? Neither of these two experiments allows us to answer these important questions.

B: The Perceiver-Experienced Self-fulfilling Prophecy.

A close examination of the information available to the perceiver is also needed in order to specify under exactly which circumstances a self-fulfilling prophecy has occurred. In this section Merton's definition of self-fulfilling prophecies will be shown to be inadequate for self-fulfilling prophecies in person perception. A new definition will be proposed that is more suitable, and the examples will be evaluated in the light of this new definition.

The problem with Merton's definition is that it is primarily concerned with the type of self-fulfilling prophecy that results in an enduring change in the target of the prophecy or expectation (eg. the bank that goes broke or the child whose education accelerates). However in person perception, where the primary interest is in changes in the perceiver's representation of the target, phrases like "false definition" and "come true" are less useful.

Consider the types of prophecies or expectancies held in the the experiments described here. The expectancies could be phrased in two ways, in terms of behaviours or dispositions. In Snyder et al's case this would be the difference between expecting the allegedly attractive females to *behave* in a friendlier way, or expecting them to *be nicer people.* While this may seem to be a trivial distinction it is crucial to Merton's definition. If the male perceivers initially expected and then concluded that the attractive females were nicer people, then they were clearly wrong in their conclusion since there was no difference in the stable dispositions of the females in the two randomly allocated groups. If, however, the male perceivers expected and concluded that the attractive females would behave in a nicer manner, then they were correct (even if for the wrong reasons). It can be inferred from the theoretical basis of attribution theory that individuals are much more likely to think of others in terms of their dispositions (or the causes of their behaviour) rather than the raw behaviour itself (Nisbett, 1975), but this is perhaps more of an assumption than a fact, and debatable.

This point may seem clearer with the following consideration of Merton's example. If the depositors of the Last National Bank thought that the bank was about to collapse, (a prediction about the future actions of the bank) they were correct and later they were proven to be correct. If, however, they were claiming that the

bank was, before the collapse, weaker than the other banks (a "disposition" of the bank), then that statement was false, and remains false even though the bank collapsed.

These difficulties can be overcome by adopting a new definition of a self-fulfilling prophecy, that does away with this problem by addressing itself to the primary focus of interest, the perceiver's representation.

A self-fulfilling prophecy has occurred when:

1. A perceiver holds an expectation about a target

2. The perceiver interacts with the target

and

3. As a result of this interaction the perceiver holds the expectancy to be more certain than before.

This definition disregards the confusing terms "false definition" and "come true.". Instead, it focuses on the crucial aspect of expectancy confirmation as far as person perception is concerned -- the way that expectations come to be held with more rather than less certainty. In other words, it captures the essence of the self-fulfilling prophecy, the way in which beliefs, rather than gradually coming into line with reality, can become self-perpetuating.

It is worth noting at this point that it is very difficult to prove that a particular interaction

constituted a self-fulfilling prophecy. Instead one usually has to rely on the experimental designs usually used in experiments on human rationality. Perceivers are typically divided into two groups and a different expectancy is given to those two groups. They are then given a chance to interact with the targets of those expectancies. If the average expectancies of the groups is larger after the interaction rather than before the interaction, then a tendency for self-fulfilling prophecies to occur has been demonstrated.

How do the conclusions drawn from the two social psychological experiments reviewed so far change when considered in the light of this new definition?

By its very nature, the crucial question about the direction of change of the perceiver's expectation cannot be answered in Word, Zanna & Cooper's (1974) experiment. Because the interviewers used in their second investigation were confederates, their opinions of the interviewees were not measured. It would be possible, though, for a white interviewer who started out by having very different expectations of black and white job applicants to conclude that the black applicants (even if they underperformed for the reasons explored in the experiments) were not as terrible as he had expected, and thus to moderate his negative opinions of them. It would only be if he were even more certain after the experiment that the black interviewees were inferior that a self-fulfilling prophecy would have

taken place.

The data from Snyder, Tanke & Berscheid's experiment, by contrast, should provide a perfect testing ground for this new way of looking at expectancy confirmation. Did the male perceivers strengthen their (possibly tentative) guesses that the attractive females would be more sociable?

The crucial test is to see whether the males' ratings of the females in the attractive and unattractive conditions on the Impression Formation Questionnaires were more or less divergent after the conversations compared to before them. This would show as an interaction between time of testing (before verses after) and attractive verses unattractive conditions, if a two-way analysis of variance was computed. It is not enough to simply find that, after the conversations the males in the two conditions still rated the females differently. This could simply mean that they had not yet learned enough about the females, but that in time the two groups would have completely overcome their initial misperceptions.

Surprisingly and without any justification there is no mention of the males' post-interaction scores in the results section, even though it is clearly stated in the method section that the measures were taken both before and after! Personal correspondence with Mark Snyder (1985) revealed that an earlier draft of the paper did

report that there were still differences between the males' perceptions of the females on some of the scales after the interaction. In the attractive target condition the females were seen as more exciting, whereas the females in the unattractive target condition were seen as more altruistic and more kind (ps all <0.05, two tailed). Unfortunately there is no repeated measures analysis to compare the pre-and post-interaction perceptions directly, but a simple count of the number of variables showing significant differences after rather than before the interaction shows that before the interview there were four variables significant at the 0.025 level (the level of signif-icance quoted in the original report) before the interview, and only two afterwards. It seems as if the males' perceptions of the females in the two conditions were converging, not polarising.

Evidence from other studies of self-fulfilling prophecies also fails to demonstrate that the expectancies have a self-perpetuating nature. In a very similar experiment Andersen & Bem (1981) reported some measures taken after the interviews, although they were different measures and analysed seperately from the scores taken at the beginning of the experiment. A casual inspection of the results does reveal, however that the attractiveness stereotype was probably diminishing over time. Before the interaction there was a highly significant main effect for attractiveness ($p<0.00001$), but after the interaction

there was no main effect for liking, just a non-significant trend for a three-way interaction ($p<0.06$) showing that all perceivers except androgynous females ended the interviews liking the allegedly attractive targets slightly more than the allegedly unattractive ones. While bearing in mind the dangers of comparing two different dependent variables directly, it seems as if the attractiveness stereotype soon diminishes rather than perpetuates itself.

Other evidence, albeit indirect, also supports this position. Thomas & Malone (1979) used Snyder, Tanke & Berscheid's data (among several others) for a completely different purpose -- to investigate the closeness of fit of various discrete-state probabilistic models to social interactions of different sorts. Although the main content of their findings are of no direct relevance to the arguments presented here, there is one incidental finding of great interest.

Thomas & Malone employed their own rater-judges to rate each utterance from Snyder et al's subjects over the first four minutes of the conversation. They used a six point scale of animation from 1, very flat and expressionless, to 6, full of energy and excitement. If the males were becoming slowly more convinced that the attractive females really did live up to their stereotypes, then one would predict that the males and females from the two conditions would start the conversation at the same level of animation and slowly diverge, the

males and females in the attractive condition slowly becoming more animated compared to the males and females in the unattractive condition.

The results show exactly the opposite. The males in the attractive condition were already more animated in their first utterance, and this difference between the groups remained fairly constant over the rest of the four minutes analysed. The females had already responded to this differential treatment by their second utterance, and the difference between the two groups showed, if anything, a slight decline from there onwards throughout the conversation (see Thomas & Malone, Figure 1, p.347).

When analysed in this way, far from looking like self-fulfilling prophecies, these effects now appear to be more similar to another effect known to social psychologists. An experiment conducted by Argyle & McHenry in 1971 investigated the phenomenon whereby the wearing of spectacles makes an individual appear more intelligent to others. They found that while the effect works well when the stimuli are presented only as photographs, any longer exposure using video-recordings caused the effect to disappear completely. The evidence from the self-fulfilling prophecy literature, in as much as it can provide evidence to these answers at all, seems to suggest that expectations, beliefs and prophecies are moderated by, not enhanced by, social interaction. What has been proved in the two experi-

ments is that individuals change their style of interaction depending on their expectations about the person they are interacting with, and that this can in turn effect the behaviour of that person. Unfortunately, none of the other experiments concerning self-fulfilling prophecies in person perception have demonstrated this crucial effect of the perceiver's expectancy on the interaction and then, most importantly, on the perceiver's final representation of the target either.

Kelly (1950) only measured the students' representation of the new lecturer after the lecture. It is thus possible that the students already had formed their opinions of the lecturer from the introduction, and they were reduced slightly by the interaction. It is also, of course, difficult to generalise from a lecture to any other social interaction.

Zanna & Pack (1975) were only concerned with the self-presentation of targets (females) depending on the expected characteristics of the perceivers. It was found that females presented themselves in a manner more consistent with the traditional female stereotype when they expected their attitudes and behaviour to be observed by a desirable male whose "ideal woman" was of the traditional type. There was no interaction with these males; they did not exist.

A further type of a demonstration has used the impoverished environment of the prisoner's dilemma game

as the form of interaction between perceiver and target (Kelly & Stahelski, 1970; Snyder & Swann, 1978b). This in itself probably makes any generalisation to issues outside of that particular prisoner's dilemma game unjustified (Morley & Stephenson, 1977). Snyder & Swann's main dependent variable was the tactics the subjects used in a reaction time game with and without noise interference. It is likely that the subjects were simply familiarising themselves with the rules of a rather bizarre game unrelated to the processes of social interaction. Again it is impossible to say whether the "*reign of error*" (Merton, 1957) would continue to perpetuate after the initial manipulation, or would fizzle out.

Kelley & Stahelski's study of the "Triangle Hypothesis" is, perhaps, the empirical work that comes closest to demonstrating how self-fulfilling prophecies work in the real world. It is not, however, a true experiment. They did not allocate subjects to groups randomly and manipulate their expectancies. Instead they relied on two different stable types of individuals, cooperators and competitors (as measured by asking subjects what they perceive as the goal of the prisoner's dilemma game to be; to cooperate or to compete). They demonstrate how their styles of interaction could lead to a perpetuation of the differences in the world views of these two groups, which could in turn lead to a re-inforcing of their world views (and so on). More specifically, it was found that competitors would fail

to recognise cooperative acts by others, and force them to act competitively, reinforcing their world view that the world is made up of only competitive, not cooperative individuals. This is not the same, however, as actually creating and studying the phenomenon experimentally. These personality types contain complex constellations of cognitive and motivational factors, and it is impossible to fully understand the process by simply observing their behaviour under different conditions.

Other Self-fulfilling prophecies have been concerned with attitudes about political issues rather than perceptions of people (Lord, Ross & Lepper, 1979), or concerned with self-perception rather than other the perception of others (Swann & Hill, 1982b; Snyder & Skrypnek, 1981). Even Darley & Fazio (1980) could find no direct evidence of perceivers automatically interpreting a target's behaviour as confirmation of the expectancy. Instead they gloss over this point and cite research that <u>suggests</u> that this <u>might</u> be the case.

<u>Further Statistical Criticisms of Snyder, et al, 1977.</u>

As well as the main criticisms developed from the new definition of the self-fulfilling prophecy, other details of Snyder, Tanke & Berscheid's experiments are also either dubious or simply wrong. For example, one of the most impressive claims made by Snyder et al was

that the only scales that detected significant differences between the two attractiveness conditions were those scales where the males had expected there to be differences. Snyder et al arrived at this conclusion after comparing the 21 scales where the males had shown a large difference in their expectations of the attractive versus unattractive females with the 13 scales where the initial differences had been small. The judges found the differences in the means to be in the predicted direction for a large majority of the high discriminators (17/21), but for the low discriminators the proportion was barely above the chance level (8/13). Binomial tests showed that the first proportion (81%) was highly significant and the second proportion (62%) was not significant. However, no direct comparison was made between these two proportions to see whether they differed <u>from each other</u>. The obvious way of doing this would be by using a chi-square test on the 2x2 table of high discriminators vs low discriminators and predicted direction vs opposite direction. The calculations performed on this table (Table 1.2) actually reveal that the difference between these two proportions is not significant, $X^2 = 1.05$, df=1, $p > .2$, one tailed.

Snyder et al seem to have been forced to use clumsy statistical procedures because they had a very large number of dependent variables and no clear rationale for analysing them. They were probably trying to cover all possibilities by taking as many measures as possible.

This is often a useful tactic to employ -- indeed McGuire's influential paper published in 1973 on more advanced conceptual and statistical approaches strongly advocates the use of many variables and advances multivariate analyses. The dangers of this approach are, though, that experimenters can find themselves with too many variables for any concise hypothesis testing while running a very high chance of committing both *type one* and *type two* errors simultaneously. The combination of approximately three hundred dependent variables (many of those already aggregated) and few clearcut theory-driven tests led Snyder et al into confusion patched up with weak, inappropriate post-hoc statistics. There were other dubious facets of the empirical work too -- for instance 13 of the 51 pairs of subjects were rejected from the experiment, some for reasons that were obviously only adopted after the experiment -- such as the males feeling that an excessive age gap between them and their females had affected the interaction! It all points to a poor theoretical understanding of what to look for in the first place.

Summary and Conclusions.

1. The chapter starts with a description of Merton's initial conception of self-fulfilling prophecies. Then three experiments purporting to demonstrate self-fulfilling prophecies in social psychology are described.

2. Two other reviews of the literature on self-fulfilling prophecies in social psychology are discussed, and shown to contribute little to a real understanding of the phenomenon.

3. It is argued that there are two fundamentally different sorts of self-fulfilling prophecy in social psychology. The first is primarily concerned with the enduring effects of the perceiver's expectation on the target, the second is primarily concerned with the enduring effects of the perceiver's expectancies on his or her own representation of the target.

4. Furthermore, it is shown that Merton's definition of self-fulfilling prophecies is unsuitable for this second type. This is the type that is usually of most interest in the study of person perception.

5. The exact criteria that need to be met to prove the existence of a self-fulfilling prophecy are outlined in detail. It is demonstrated that the considerations are different for the perceivers who hold the expectancy and the observers who are naive to the expectancy and observe the target during the interaction. It is also argued that the role of these observers or judges has been over-rated in previous experiments, and they cannot determine whether a self-fulfilling prophecy has occurred.

6. The crucial feature of this second type of self-fulfilling prophecy is that perceivers are more certain of their expectancy after interacting with the target, either because of the way in which they elicit behaviour or interpret that behaviour from the target. While some experiments have illuminated how certain phases in the phenomena could occur, none of the experiments that claim to have demonstrated self-fulfilling prophecies in person perception have, in fact, found this polarising effect on the perceiver's expectation. What evidence there is suggests that exactly the opposite is going on -- rather than being made more extreme, expectancies are moderated by social interaction.

7. Because of these factors, and also other statistical and methodological faults in these experiments, it is concluded that self-fulfilling prophecies have been poorly understood in social psychology, particularly in the field of person perception. While they may well occur, a convincing experimental demonstration has yet to be carried out.

Table 1.1. Snyder, Tanke & Berscheid's (1977) data allegedly demonstrating the selective confirmation of only the "*High-Discriminator*" traits.

	Difference in Predicted Direction? Yes	No	
High Discriminators	17	4	21
Low Discriminators	8	5	13
	25	9	34

$X^2 = 1.05$, ns., one tailed.

Chapter 2
Hypothesis Testing in Social Interaction.

Introduction

In chapter 1 the literature on self-fulfilling prophecies in person perception was described in depth and critically reviewed. The rest of the thesis will deal with a closely related phenomenon, the self-confirming hypothesis. The present chapter will start by defining the difference between the two phenomena. It will then go on to give a detailed description of the existing literature on hypothesis testing in social interaction. This work attracted much acclaim at the time that it was published, and continues to do so. The reasons for this are explored by considering the climate of reseach in social psychology at the time the research was published. It is concluded that the research fitted exactly with the other influential publications at the time, sharing the same "*model of man*" and the same roots in cognitive psychology; this was probably as responsible for the interest shown in the hypothesis testing paradigm as the intrinsic merit of the experiments themselves.

The description of the literature on the testing of hypothesis about other people in social interaction will set up some criticisms of the rationale of this work, and of the methodology used in these experiments.

Chapter 3 will contain criticisms of Snyder & Swann's experiments, and of Swann's PhD thesis which they draw upon. The work described in this chapter and the criticisms of it in the next chapter will form the basis for all of the empirical work to be presented in chapters 4 to 8 of this thesis.

Self-Confirming Hypotheses and Self-Fulfilling Prophecies.

The difference between hypothesis-testing and the phenomena described in the last chapter dealing with beliefs or expectations has hardly ever been fully articulated in the literature (exceptions are Burchell, (1984) and, to a lesser extent, Snyder (1985)), yet it will argued here that they are conceptually quite distinct. Furthermore, a failure on the part of experimenters and theorists to make this distinction has often led to false conclusions being drawn from empirical work.

Many recent social psychology text-books fail to make any distinction at all between expectations and hypotheses in this context (eg. Wegner & Vallacher, 1981, p228; Ross & Anderson, 1982, p151; Hamilton, 1981(b), p120); the two are discussed interchangably. Some experiments have been called examples of hypothesis testing by the experimenters when, in fact, they are (by the definition that will follow) really about acting on expectations (eg. Darley & Gross, 1983), and other

experiments combine the two processes without distinguishing between them (eg. Carver & de la Garza, 1982).

Snyder has performed experiments both on the effects of expectations (eg Snyder, Tanke & Berscheid, 1977; Snyder & Uranowitz, 1978) and on the effects of hypothesis testing (eg. Snyder & Swann, 1978; Snyder & Cantor, 1979). Usually he talks about the two types of experiments separately, but more recently he has said explicitly what he considered the difference to be (Snyder, 1985). Even then he says that while there may be a theoretical distinction, they were "*in practice not different*" (p.261). It will be argued in Chapter three, however, that he is still guilty of treating the two processes interchangably at times, and of making unwarranted inferences from one to the other which severely undermines the value of his work and the conclusions that he draws.

There are two fundamental difference between a hypothesis and an expectation. Firstly, there is the belief component, equivalent to the "prior odds" in a Bayesian analysis. An expectation has a definite effect on the prior odds. For example, the males in Snyder, Tanke & Berscheid's experiment expected the females to be nicer people in the "attractive target" condition. By contrast, testing a hypothesis needs no prior expectancy. A subject simply trying to find out whether a particular target posesses a particular trait does not need to have any prior assumptions about the likelyhood

of that hypothesis being true.

The second distinction between a testing a hypothesis and acting on an expectation concerns the goal of the act. When testing a hypothesis a subject is explicitly told that their task is to discover certain things about the target, for instance their level of extraversion. Subjects acting on expectations may not be involved in a quest for knowledge at all; they may be to more concerned with, for instance, managing the impression the target forms of them.

It is, of course, possible to combine the two processes in one interaction; For instance Word, Zanna & Cooper (1974) not only used the subjects pre-existing racist beliefs, they also explicitly set them the task of finding out certain things. Similarly, subjects in some of the hypothesis testing experiments were also given explicit information that led them to believe that their hypotheses were particularly likely to be either true or false (eg Snyder & Swann, 1978, investigations 1 and 3).

One gap in the literature is a knowledge of when, why and how individuals actually set themselves hypotheses to test (this point is considered again in more detail in chapter 9). It is quite plausible that hypothesis testing takes place mainly when an individual has an uncertain expectation about a target. For instance, it is possible that a subject could test a

hypothesis that is at variance with his expectations. For instance, in Word, Zanna & Cooper's first experiment an interviewer may have been testing a black target for suitability for the task to follow, whilst simultaneously believing them to be probably unsuitable because of their colour. It is, however, still important to conceptualise the two processes as seperate.

To summarise, a hypothesis is merely a statement about the nature of the world. These statements are, by implication, to be tested and their veracity determined. By contrast, expectations are predictions or diagnoses about the state of the world, and are not necessarily to be tested but simply to be used in deciding how to act in a particular situation. In Snyder's words this difference is between "*reality testing*" and "*reality coping*" (1985, p. 261).

Snyder & Swann's Key Experiment

Snyder & Swann (1978) point out that one of the functions that conversations fulfil is to satisfy the actor's quest for social knowledge. One way in which individuals may structure this quest is by testing hypotheses. An example of this may be if one wishes to find out whether an individual is representative of the stereotyped group he is a member of or whether a rumour we have heard about someone's extreme friendliness is true.

For instance I may know that a new member of staff in the department is an American (a nationality stereotyped as being very forward and sociable), or an old collegue may have told me that she was the life and soul of her old department, or I may be considering inviting her to a party but want to make sure that she is outgoing enough. I would probably wish to check on her personality myself, to see what she is really like -- particularly if I am going to have to work with her. So, according to Snyder and Swann's thinking I would have formed the hypothesis that the new lecturer was an extravert, and be preparing to test that hypothesis.

Having obtained a hypothesis to test, the hypothesis tester will need to collect data with which to test the hypothesis. Snyder and Swann identify three distinct strategies that could be used in the gathering of data. Firstly, the search could be geared towards

finding evidence that would support or confirm the hypothesis. Thus, for instance, a test of an extravert hypothesis would involve gathering as much evidence of sociability and outgoing behaviour as possible. Alternatively, the hypothesis could be tested by attempting to gather evidence that would disconfirm or weaken the hypothesis. Using this strategy an individual might attempt to search for examples of shyness or social ineptness in testing an extravert hypothesis. If such evidence were found, then one could conclude that the hypothesis was wrong or inaccurate. The other possible strategy would be to devote an approximately equal am ount of time to the search for both hypothesis-confirming and hypothesis-disconfirming evidence.

The simplest method of searching for information in a conversation is by asking questions. Thus, in order to find out how people test hypotheses in social interactions, Snyder and Swann gave individuals hypotheses to test and recorded the sort of questions that they asked.

The actual procedure that Snyder & Swann (1978, investigation 1) used was as follows. An individual (called "*the interviewer*") was told that they were taking part in an experiment to find out how people come to understand each other. They were to ask questions of another subject waiting in another room in order to find out about the personality of that person (known

hereafter as "the *target*"). More specifically, they were to find out whether the target was the type of person described on a card they were given. They were then told that the profile they were given on the card dealt in *abstract generalities and global characteristics*, but getting to know someone involves finding out concrete information and specific facts about what the person actually thinks, feels and does. They were thus told to select 12 questions to ask of the target, to find out whether the target's *specific beliefs, attitudes and actions in life situations matched the general characteristics described in the profile.*

They were given a list of 26 questions to choose from by the experimenter. They were told that this list was typical of "*topic areas often covered by interviewers*" and were instructed to select the 12 questions that would provide the best information with which to test their hypothesis about the target.

When the participants had chosen their 12 questions to ask they were told that there was, in fact, no target and the interview would not take place.

There were two orthogonal manipulations in the experiment: The hypothesis they were given to test and the certainty of the hypothesis. The hypothesis was manipulated by the personality profile that the participants were given. Those assigned at random to the "*extravert hypothesis condition*" were given the

following description:

> *"Extraverts are typically outgoing, sociable, energetic, confident, talkative and enthusiastic. Generally confident and relaxed in social situations, this type of person rarely has trouble making conversation with others. This type of person makes friends quickly and easily and is usually able to make a favourable impression with others. This type of person is usually seen by others as characteristically warm and friendly."*
>
> (p. 1023)

The participants in the "Introvert Hypothesis Condition" were given a description of a typical introvert thus:

> *"Introverts are typically shy, timid, reserved, quiet, distant, and retiring. Usually this type of person would prefer to be alone reading a book or have a long serious discussion with a close friend rather than to go to a loud party or other large social gathering. Often this type of person seems awkward or ill-at-ease in social situations, and consequently is not adept at making good first impressions. This type of person is usually seen by others as characteristically cool and aloof."*
>
> (p. 1024)

The certainty of the hypothesis was manipulated by giving different subjects different reasons as to the origins of the personality profiles. In the "High Certainty" condition participants were told that the profiles were generated from a personality test taken by the target the week before, and their task was to see whether it was an accurate portrayal of the way the subjects actually came across in real-life situations. A pre-test had confirmed that the subjects had a lot of faith in personality tests.

The subjects in the "Low Certainty" condition were given no information to lead them to believe that the hypothesis would be correct. They were simply told that it was a description of a "*type of person familiar to us all*" (p.1204) and they had to see how well it described the person they were to interview.

The dependent variable was the frequency with which various types of questions were chosen by the participants. The list of questions had been classified into three groups by nine rater-judges. The categories provided by the experimenter for this classification task were:

1. *Extravert Questions*. These are questions that one would ask of someone already known to be an extravert. An example is "*What kind of situation do you seek out if you want to meet new people?*".

2. *Introvert Questions.* These are questions judged to be ones that one would ask of someone already known to be an introvert. "*What factors make it hard for you to really open up to people?*" is an example of this category.

3. *Neutral Questions.* This category was made up of questions irrelevant to the introversion-extraversion dimension or questions for which there was no consensus among the rater-judges. "*What are your career goals?*" or "*What do you think the good and bad points of acting open and friendly are?*" are both examples of this category.

The results were very clear-cut. Participants chose more extravert questions to ask when testing the extravert hypothesis than when testing the introvert hypothesis. The opposite effect occurred with the introvert questions -- these questions were chosen more frequently in testing the introvert hypothesis than the extravert hypothesis. The neutral questions were chosen with an equal but low frequency in both hypothesis conditions.

The choice of questions was, however, completely uneffected by the certainty manipulation, there being no main effect for Certainty and no *Certainty by Hypothesis* interaction.

This led Snyder and Swann to their first major conclusion -- that when testing hypotheses individuals search preferentially for evidence that is supportive of the hypothesis. This phenomenon they termed the "*Confirmatory Bias*".

The concept of such a confirmatory bias is not new to science, philosophy or psychology. Studies of hypothesis testing in non-social settings have shown that individuals will consistently search for confirmation of a hypothesis in a numeric reasoning task (Wason, 1960) or a logical reasoning task (Wason & Johnson-Laird, 1965). The philosopher Francis Bacon drew attention to the failing of human reasoning that leads it to find support for, rather than evidence against, any theory (1620). More recently philosophers of science have warned scientists of their tendancy to be over-zealous in their search for support for their hypotheses and thus retain those hypotheses longer than is appropriate, causing inefficiancy, conservatism and delay in the advance of knowledge (Popper, 1959).

But the finding that people search for confirmation of their hypotheses does not lead to any important conclusions in itself. If the finding is to be of anything more than academic importance, the crucial question is whether this search for confirmation actually leads people to inaccuracies or biases in their inferences about the truth of the hypothesis. In other words, are the impressions people form about others a

function of the particular hypothesis they happened to be testing? In the second of the four investigations in their 1978 paper, Snyder & Swann attempted to find an answer to this question. How would this *confirmatory bias* in the choice of questions to ask affect the opinions that were formed about the likely introversion or extraversion of targets after they had been asked the questions selected by the interviewers?

The Effects of the "Confirmatory Bias"

The procedure for Snyder & Swann's second investigation was essentially the same as had been used in investigation 1, but the interviewers were allowed to continue after they had selected the questions and ask them of a second subject, the "*target*". Since the certainty of the hypothesis was seen to have no effect on the choice of questions, this manipulation was dropped and the subjects were all run under the "low certainty" condition; that is to say they were given no reason to suspect that the hypothesis was any more likely to be accurate than by inaccurate.

The targets were told that they were to be interviewed by another student, and were instructed to answer all of the questions in "*as informative, open and candid a manner as possible.*" (p. 1206). When the interviewer had selected 12 questions to ask she addressed these questions to the target via a microphone and headphone intercom. The entire conversation was tape-recorded on a stereophonic system so the interviewer's and target's channel could be played back seperately.

As before, the types of questions selected for the interview were monitored. Analysis showed that the "confirmatory bias" had been faithfully replicated; Introvert questions were more likely to be selected to

test the introvert hypothesis than the extravert hypothesis, and extravert questions were more likely to be employed to test the extravert hypothesis than the introvert hypothesis.

The more important questions involved the way that targets responded to these questions, and whether the interviewers regarded the behaviour of the targets as being evidence supportive of their initial hypotheses.

To determine whether the targets had actually come to behave in a manner consistent with the hypotheses being tested by the interviewers, "naive" judges listened to the recordings of the targets' responses to the questions asked during the interview, and rated the replies on a number of scales such as talkative-quiet and introverted-extraverted. It should be remembered that neither the targets nor the "naive judges" were aware of the hypothesis being tested by the subjects in the role of interviewer.

And could the judges detect any differences between the targets who were being tested for extraversion and those whose interviewers were testing for introversion? The results showed that this was clearly the case. On a variety of measures the targets in the *extravert hypothesis* condition were perceived to be reliably more confident, poised, energetic and extraverted than the targets in the *introvert hypothesis* condition.

The pattern of events that may occur when individuals use social interactions to test hypotheses is, according to Snyder & Swann, becoming clear. It all starts with the questions that they prepare to ask to test their hypotheses with. They are likely to preferentially choose those questions that probe for evidence that will be supportive of their hypotheses. When asked these questions the targets will reply in a way that, more often than not, is consistent with the hypothesis. Thus, merely by testing a hypothesis about someone (even though they are not aware of this hypothesis, and even though the interviewer has no reason to believes that the hypothesis is accurate) one is likely to make them behave in a manner that is consistent with the hypothesis.

The search for the limits of the Confirmatory Bias

Snyder & Swann's next step was to try to find the limits of the phenomenon, or how easy it is to lessen or remove the confirmatory bias. Their first attempt involved looking at situations where it was obvious to the hypothesis testers that their hypotheses were likely to be false. They did this by incorporating more information for the subjects in the instructions given before the selection of questions.

The manipulation in the third investigation involved base rates. All of the subjects were given the extravert hypothesis to test, but half of them were told that the target had been selected from a sophomore (women's college society) that had all been given personality test and it was found that 23 of the 30 women were extraverts. In the other condition the "interviewers" were told that only seven out of the 30 were found to be extraverts. In order to ensure that the interviewers fully appreciated the implications of this manipulation of prior odds (given the consistent finding that base rates are often ignored when making inferences about others, eg. Tversky & Kahneman, 1982) the high or low chance of the hypothesis proving to be true in the two conditions was emphasised. A manipulation check verified that the subjects were sensitive to the high or low probability of the hypothesis being accurate.

However, the choice of questions seemed to be independent of the likelihood of the hypothesis being accurate. Snyder & Swann reported that the confirmatory bias was still present and just as strong as in the extravert hypothesis conditions in the previous investigations, regardless of whether the participants believed that their hypothesis was likely to be true or false. It seems from this that it is the very fact of testing a hypothesis that induces the search for confirmatory evidence, not the likelihood of that hypothesis being true. Hypotheses can still be self-confirming regardless of what the tester perceives the prior odds to be.

In their final investigation (investigation 4, 1978) Snyder & Swann tried yet another ploy to reduce or eliminate the confirmatory bias -- a motivational intervention. The subjects were told that a prize of $25 would be awarded to the person who selects the best questions for finding out about the target. Again there was no evidence that this diminished, let alone overrode the confirmatory bias.

This quest for the limits of the confirmatory bias was to be the aim of two more journal papers by Snyder and his collegues. In 1980 Snyder & Campbell tried yet another approach. They argued that one of the reasons that subjects showed such a robust and persistent propensity to search for confirmatory evidence when

testing hypotheses was because the hypotheses themselves were phrased in terms of what information would lead to their confirmation, but no mention was made of the sorts of information that would lead to their rejection. Snyder & Campbell ran some subjects (the control condition) as usual, but in the experimental condition the personality profiles were changed so they contained not only examples of what an extraverts (or introverts) were, but also instances of what they were not. For example the extrovert condition now contained phrases such as "*... are rarely shy, timid, reserved, quiet, distant and retiring*" (p.423).

Did this increase in the availability of the instances whose presence would be likely to disconfirm the hypothesis lead to a weakening of the tendency to search preferentially for data that would confirm the hypothesis? Again the answer was no; there was no difference between the subjects who had hypotheses framed in both positive and negative terms and those whose hypotheses were framed only in terms of positive, hypothesis-comfirming attributes. In fact in another, unpublished paper Snyder & White (1978) went one stage further and phrased the hypotheses exclusively in terms of disconfirming attributes. This still led to an undiminished effect whereby questions were selected to preferentially seek out confirmation of the hypothesis. While these results suggest that the confirmatory bias is not simply a function of the mental representation of hypotheses in terms of positive attributes, Snyder &

Campbell do consider that there may be another explanation; perhaps the manipulation of the personality profiles given to the subjects may simply not have been powerful enough to overcome the very strong intuitive notions of extraversion and introversion known to virtually everyone in our culture (especially university students who were the experimental subjects). Thus a better test of this theory that the framing of the hypothesis is important would have to be conducted using a different trait less well popularised in lay terms.

Testing Hypotheses about Social Stereotypes.

Snyder's quest for a tactic that would prevent individuals from gearing their search towards the confirmation of hypotheses continued into a slighty different paradigm; he turned to investigating the way in which individuals test social stereotypes (Snyder, Campbell & Preston, 1982). The procedure was essentially the same, looking at the effect of opposing hypotheses on the types of questions selected from a list. This time various "educational interventions" were tried to see if they could weaken or eliminate the confirmatory bias. Subjects read one of four short passages before selecting their questions. These passages gave advice on how to go about testing social stereotypes most effectively in order to sensitize subjects to the value of disconfirming evidence. The advice ranged in its degree of explicitness from one which merely pointed out that our assessment of the accuracy of a hypothesis should decrease if information not consistent with the hypothesis is found, to one which suggested selecting questions which would uncover the way in which people do <u>not</u> fit their stereotypes.

When these four conditions were compared to a control condition without any intervention the predisposition to select questions that would tend to elicit confirmatory evidence was just as strong. A simple educational intervention was evidently not enough

to overcome the confirmatory bias.

In a second part of the study yet another intervention was tried. This time it was pointed out to the participants that while they were trying to find out about the target, the target would be trying to form an impression of them, and that impression would depend on the questions that they chose to ask. Thus the subjects were told that they may appear to be closed-minded if they ask only questions that do not give the target the chance to show ways in which he does not fit the stereotype.

Finally Snyder, Campbell & Preston have identified a method of overcoming the confirmatory bias! In this last condition the subjects asked the same sorts of questions regardless of the hypothesis they were testing, either the warm stereotype of a counsellor or the cold stereotype of a researcher. Snyder, Campbell & Preston go on to explain why this "*impression management*" intervention succeeded in combatting the confirmatory bias when all others have failed. They suggest that any intervention that implicates one's own self and the opinion that others have of one is likely to bring about a high degree of involvment and thus will be considered a lot more carefully by the participants than the impersonal educational interventions.

Historical Hypothesis Testing

It is worth noting at this point that as well as using social interactions to seek out new information in order to test hypotheses, it is also possible to test hypotheses using information available to us in other ways. An investigation of hypothesis testing using information already stored in one's memory was the subject of another of Snyder's experiments (Snyder & Cantor, 1979). The findings resemble those reported on hypothesis testing in social interaction. Individuals preferentially recalled data that was supportive of a hypothesis, and under-reported information that was inconsistent with the hypothesis. This again led the hypotheses to be self-confirming.

This paradigm of "historical hypothesis testing" is, however, sufficiently different from hypothesis testing in social interaction to be dealt with seperately later in the thesis. It will be described in more detail in chapter 8, where the results of an experiment on historical hypothesis testing are reported.

Practical Considerations.

So, what conclusions can be drawn from this series of experiments? Snyder, (1981, Snyder & Gangestead, 1981) outlines what he considers to be the practical implications of this research programme. He says that the mechanisms of some of the best-known phenomena in social psychology can best be understood in terms of his research. For instance, the stability with which erroneous social stereotypes persist over time can be explained in terms of the confirmatory bias. Even though an individual may attempt to find out for herself whether, say, researchers are really cold, unemotional people, she is likely to test this by asking questions that would make anyone respond in a cold, unemotional way and conclude, on the basis of her own personally aquired evidence, that the initial hypothesis was indeed valid.

Another way in which this manner of testing hypotheses may handicap us is in the testing of our hypotheses about the nature of the world. Snyder gives the example of the way in which psychiatrists test their theories. If, for example, a psychiatrist adheres to a theory that puts a lot of emphasis on relationships early in life as the causes of later neuroses, she is likely to probe deeply into the early relationships of all of her patients until they eventually disclose a potential expla nation. At this she is likely to stop her search and be that little bit more certain that her

theory of early relationships is correct. What she would be unaware of was that it was her confirmatory strategy that lead her to the confirming evidence -- if she had enquired about problems in the early relationships of people who did not have psychiatric problems the chances are that she would also have found a rich supply of possible roots of a patient's neurosis.

As a general rule, the reason why hypotheses concerning the social world are likely to be so reactive to hypothesis testing strategies is because of the rich variety of behaviours indulged in by any individual. An argument central to modern Social Psychology and personality theories is that there is little consistency in behaviour over situations (Mischel, 1968); most people will be strongly influenced by situations and their behaviour will change accordingly. While the extent of this lack of consistency is still debated, there is a general consensus in social psychology that most individuals are capable of behaving in very different ways in different situations. Thus most targets will be able to give plenty of instances of either characteristically extraverted or characteristically introverted behaviour if asked to do so. Similarly, the backgrounds of a group of men who were selected on the basis of them leading healthy, normal adult lives were found to be rife with examples of traumatic events and "pathogenic factors" that, if they were to complain of psychiatric symptoms, would almost certainly be evoked in the expla nation of the ontogeny

of those symptoms (Renaud & Estess, 1961).

This being the case Snyder argues that no hypothesis testing strategy is likely to be any more accurate than any other. If interviewers used the disconfirmatory strategy they would cause the targets to behave in a manner inconsistent with the hypothesis, and hypotheses would become self-disconfirming hypotheses. Snyder says that even an "equal opportunity" strategy would still be reactive, and produce "half and half" people (1981, p. 301). Because of the very nature of the social world (unlike the physical world) any attempt to test a hypothesis is going to affect the validity of the hypothesis. Perhaps, Snyder pessimistically concludes, the only way to test a hypothesis faithfully in the social world is through a sleuth-like following of the target, so he can be observed in a wide variety of situations which could be somehow averaged by the observer/hypothesis tester. Since it is unlikely that we will take such a complex approach to our gathering of social knowledge, Snyder (1981) ends by stating that people "*create a world in which hypotheses become self-confirming and beliefs become self-perpetuating.*" (p.301).

The Climate of research in the late 1970's

Snyder's research on the testing of hypotheses in social interaction was received with great interest at the time it was published, and it continues to be cited widely not only in mainstream social psychology but also in applied fields such as personnel selection (Sackett, 1982), clinical psychology (Witkins, 1982)and management training and decison theory (Keisler & Sproull, 1982). To demonstrate the way in which this work has been accepted into the received wisdom of Social Psychology one needs only to look at the number of the citations the research has obtained. In the 1979 edition of the *Social Science Citation index* there were three journal articles that cited Snyder's hypothesis testing research. This rose to five in 1980, 37 in 1981, 42 in 1982, 32 in 1983, 52 in 1984 and 55 in the 1985 edition (see figure 2.1). In addition to these, Snyder's hypothesis testing research is now cited in most of the general textbooks written on Social Psychology. Snyder has also been invited to write chapters detailing this work in books on attribution theory (Harvey, Ickes & Kidd, 1981) stereotyping (Hamilton, 1981b) and experimental social psychology (Berkowitz, 1985). Being included in these prestigeous books is further evidence of the acclaim his work achieved among his peers. A further compliment to Snyder comes from Pettigrew who remarked that Snyder's presentations to the American Psychological Association's conferences were " .... *characterised by standing-room only crowds.*" (P.303,

1981).

It will be argued here that in order to understand this widespread acceptance of Snyder's work on hypothesis testing, one needs to look beyond just the research and findings themselves and consider the climate of Social Psychology at the time it was published. For a number of inter-related reasons Snyder's work was destined to be successful in 1978 in a way that it probably would not have been in 1968 or 1988. It fitted in with the model of man that was fashionable at the time for cognitive social psychologists, used the perspectives on emotion and motivation that were prevalent in the 1970's, made a substantial contribution to applied areas of psychology that were attracting a lot of interest at that time, drew upon the "right" sources, and complemented the other key research of its time.

Models of Man.

It has long been argued that at different times psychological research has employed different *"models of man"* (Chapman & Jones, 1980). These entail different perspectives on the purpose and mechanisms of the human mind. It is widely accepted that the progression from one model to another does not follow a strictly rational advancement according to a simple Popperian conception of scientific progress, but that there are other factors to account for these paradigm

shifts (Kuhn, 1962) or, perhaps more appropriate to psychology, *"metaphysical cores not open to falsification"* (Lakatos, 1970). These "other factors" may include influences such as the types of problems that researchers or their sponsors are trying to solve, the technology available as tools for research, new philosophical perspectives, political influences, and sometimes the process may even take on an almost random element, changing for the sake of change (see the book edited by Chapman & Jones (1980) to see how central yet controversial the role played by models in contemporary Psychology is).

The dominant models of man that influenced Social psychology in the latter half of the 1970's considered man to be an intuitive psychologist (Ross, 1977), an intuitive scientist (Nisbett & Ross, 1980; Kelly, 1955) or an intuitive statistician (Kahnemann, Slovic & Tversky, 1982; Edwards, 1965). These models are not mutually exclusive -- on the contrary it is very difficult to distinguish between them at times. They all stress that man is an information processor, they all show him as being in a quest for a greater understanding of his environment and they all emphasise the fact that he is fallible and often falls short of the cognitive processes needed to fulfil his goals in an optimal manner. This is exactly the picture that Snyder & Swann portray -- people as keen but faulty seekers of information to increase their understanding of the other members of the social world.

Motivational Considerations.

One of the common themes running through all of the research discussed so far is that it attempts to explain all of human judgement, and the shortcomings in that judgment, in purely cognitive terms. This is in sharp contrast to more psychoanalytically orientated work which explains *"irrational"* behaviour in motivational terms. None of the theorists involved in these paradigms would be willing to admit that items were forgotten because they were threatening to the individual's *ego,* or an outgroup was evaluated negatively to *project* an individual's own faults onto that group. The *"Lay psychologist"* also resorts to motivational explanations of the behaviour of others, as the proverbs *"None so deaf as those who don't want to hear"* or *"People see what they want to see"* reveal. The rejection of motivational explanations by social psychologists at the time that Snyder & Swann published their paper has to be understood by looking at the history of motivational verses cognitive explanations.

Many experiments have claimed to provide demonstrations of instances where individuals have distorted their perceptions in order to defend their own *"egos"* or boost their own self-esteem. These experiments have typically employed one of two forms, either looking at interpersonal influence or team achievement.

In the first of these types of experiments (the first and most notable being reported by Johnson, Fiegenbaum & Weiby in 1964) teachers are typically given a particular task to teach to two pupils. The experimenters manipulate the outcome such that one pupil succeeds in learning the task and the other fails. The teacher is then asked to explain the reasons for the success and failure of two pupils respectively. Researchers typically found that the teachers saw success as evidence of their capable teaching, but failure was explained in terms of the inadequacies of the pupil.

Research into the perceived causes of success or failure of groups on achievement tasks also arrived at similar findings. Subjects typically perform a task as a group, and the feedback they receive about the quality of their results is manipulated. If a group is told that they have been successful then the participants each overestimate their own contribution, but if the participants were told that the group has performed poorly then the others in the group are attributed with being more influential in causing that outcome (For a more complete review of the literature in this field see Miller & Ross, 1975). The conclusions reached by this line of research are no longer accepted, for three main reasons.

Firstly, The results of these experiments can be explained entirely in terms of inferences made on the evidence available to the subjects in these experiments, and from their experiences prior to the experiment. For instance in the teaching senarios, the teachers see one child change (learn the task) and the other child stay at the same level of understanding (not learn the task) throughout the experiment. It is therefore not surprising that the teacher attributes responsibility for the change to himself or herself, but not the lack of change (this follows logically from Mill's laws of causation).

Secondly, by manipulating the experiment, some researchers have been able to invert the normal outcome of these experiments so subjects make more internal attributions for failure than success (Ross, Bierbrauer & Polly, 1974). If these results were to be explained in terms of ego-involvement, it would require the postulation of a *"counter-defensive bias"* as opposed to a defensive bias, which is considered to be rather implausible.

Finally, there are conceptual inconsistencies in an information-processing account of these "New Look" explanations (as these ego-protective phenomena were called in the 1940's and 1950's) (Erdelyi, 1974). There is, for instance, a logical paradox in perceiving something so it can be hidden from perceptions, and the idea of a homunculus inside our heads controlling what

we are fit to see or realise is at odds with present conceptions of the mind. It is also difficult to see how a Darwinian process of evolution would have produced a species that would fail to see those items most threatening to its survival -- as Nisbett & Ross (1980) point out the purpose of such a mechanism is rather bewildering!

Given this background of research, empirical demonstrations of cognitive processes that could explain phenomena previously assumed to be motivational biases were received with great relish in the late 1970's. One of the most influential of these was Michael Ross's surveys that showed that many common marital disagreements (over, for instance, the two partners' claims about the size of their contribution to the housework) were caused by differential availability of evidence rather than by ego-protective biases (Ross, 1981; Ross & Sicoly, 1979). Similarly, Kelley & Stahelski (1970) demonstrated how competitive individuals would cause others to be competitive, and thus perceive their *"Dog eat dog"* philosophy of life to be normal, and Ross, Greene & House (1977) set forth the way in which individuals believe their own ideas and actions to be more widespread than is actually the case, the "False consensus" effect.

Thus, it can be seen that Snyder & Swann's model of man as a hypothesis tester again fits in nicely with this trend. The unwillingness of people to change from

their accepted beliefs was caused by shortcomings in their information search, rather than through some innate stubbornness or fear of change. The question of motivation was not tackled explicitly, but there was the implicit assumption that runs through much of cognitive psychology that the individual is simply motivated to achieve an increased understanding of his or her social world -- to "predict and control" (Kelly, 1955, p.4).

Applied Aspects -- Stereotyping.

There were a lot of publications in the late 1970's that addressed one of the traditional subject areas of social psychology, stereotypes and intergroup behaviour, but from a new cognitive basis. Following the Chapmans' work on illusory correlations (Chapman & Chapman, 1967 and 1969), Hamilton & Gifford (1976) and Hamilton & Rose (1980) demonstrated how simple knowledge of group membership or simple expectations of group members respectively could lead to errors in inference and stereotyping of minority groups or firmer stereotypes from mere expectations. Similarly, Taylor showed how being a solo female in an all-male group, or a solo black in an all-white group could cause observers to perceive that individual to be more prominent and more characteristic of their ethnic group -- the phenomenon even occurred when one member of a group dressed differently from the others (Taylor, 1981; Taylor & Fiske 1978, Taylor et al 1978). Rothbart conducted several experiments to show how categorisation could

effect memory processes and, under some circumstances, bolster stereotypes of groups (Rothbart, 1981; Rothbart, Evans & Fulero, 1979). The combination of Snyder's work on hypothesis testing and his work on self-fulfilling prophecies fitted nicely into this framework, leading to his 1981 chapter entitled "On the self-perpetuating nature of social stereotypes". Taken together, the joint impact of these contributions was seen as a major contribution to the understanding of stereotypes from a cognitive perspective (Hamilton, 1981c; Petigrew, 1981).

The Influence from Cognitive Psychology.

Another feature of this work on hypothesis testing was the way it drew upon the findings of "pure" cognitive psychology. As well as having the general advantages gained whenever the knowledge from one area of psychology cross-fertilise another, the application of cognitive psychology is held in high esteem at the moment, and was probably held in even higher esteem in in the late 1970's before it received criticisms for its lack of social content (eg. Tajfel, 1981; Harre, 1981). Not only was the switch to an emphasis on internal processes probably a genuine advantage for Social Psychology, it was also seen as one of the ways in which Social Psychology could overcome its latest *crisis* by adopting a more empirical and theory-driven approach. This trend was not limited to Social Psychology, but was also prevalent in personality, developmental and clinical branches of psychology as behaviourist orientations gave way to a cognitive outlook (Forgas, 1981). The application of artificial intelligence to cognitive social psychology (eg Shank & Ableson, 1977) was perhaps yet another reason why contributions from cognitive psychology should have received such a warm welcome.

Snyder drew upon three main areas of cognitive psychology to help in the generation of his experiments and the explaination of the confirmatory bias. These three fields were covariation detection, hypothesis-testing and the use of negative instances.

A: Covariation Detection

The cognitive research in the 1960's generally concluded that the lay-person was a competent *"intuitive statistician"* (Peterson & Beach, 1967). One exception to this was their inability to judge covariation from two by two contingency tables first noted by Smedslund (1963), and Jenkins & Ward (1965), and recently replicated by Jennings, Amabile & Ross (1982). In these experiments individuals are presented with information in the form of such a table, and are asked to judge whether the two variables are related. For instance, they could be given the outcome of a disease (cured vs not cured) and the medication given (drug vs no drug) and asked whether the drug increased the chance of a cure. One of the pervasive findings of these experiments is that individuals rely almost exclusively on the "present-present" cell (in this example drug given and disease cured), which Snyder & Swann compare to their subjects tendency to look only for evidence of extraversion when testing an extravert hypothesis, or introversion when testing an introvert hypothesis.

B: Hypothesis Testing.

In 1962 Wason conducted a well known experiment to explore the way individuals tested hypotheses. The hypotheses they tested centred around series of numbers

(ie. 2,4,6...) for which subjects had to guess the rules governing the sequence. Wason's emphatic (but controversial) conclusion was that subjects were very poor at testing competing hypotheses and simply continued to try and search for more and more evidence supportive of their best or latest guess. This led subjects to be exceedingly slow in discovering even a very simple rule. Later work with Johnson-Laird (1972) arrived at a similar conclusion, that people are poor at testing rules that require them to attempt falsification strategies. Again Snyder & Swann drew upon this work and applied it directly to a social setting where they found similar results.

C: Negative Instances

Finally, Snyder & Swann related their finding to research on concept formation and utilisation that concludes that individuals prefer to use positive instances than negative instances in their thinking (Hov land & Weiss, 1953). A related finding that Snyder & Swann also cite is that confirming instances have more inpact on inferences than negative instances (Gollob, Rossman & Abelson, 1973), although some recent experiments have found the exact opposite of this (eg. Hastie & Kumar, 1979) making the interpretation of earlier findings questionable.

These three related findings from cognitive psychology led to Snyder & Swann's conclusion that *"the*

structure and process of human thought fosters and promotes the ready and willing adoption of confirmatory strategies for hypothesis testing." (1978, p.1210).

Other Influential Research.

A common theme running through much of the social psychological literature since the early 1970's is the study of human rationality -- the "*validity in deductive or probabilistic reasoning*" (Cohen, 1981). Generally speaking most of this research has taken one of two forms. In the former, subjects are given a problem for which there is an answer deductable from normative criteria. If subjects do not arrive at this right answer, then they can be said to be irrational in their thought processes, whether through an inability to apply rules correctly or through an ignorance of the appropriate rules. For instance, Wason & Johnson-Laird's (1965) deductive reasoning tasks used this principle to demonstrate that individuals do not understand the logic of conditional statements properly.

The other way that is commonly used in social psychology to demonstrate shortcomings in human judgment is to give two groups of subjects logically equivalent forms of a problem, but to manipulate some other variable which either should or should not alter the logical form of the judgemental task. If subjects' responses are swayed by this other factor but ought not to be (for example their seating position when observing

a conversation (Storms, 1973) or the salience of one category of items over another (Tversky & Kahnemann, 1973)) or if they are not effected by factors which <u>are</u> relevant (for example, base rates (Kahnemann & Tversky, 1972) or consensus information (McArthur, 1972)) then shortcomings in their cognitive processes can be inferred. This research has lead to a model of man as having a limited capacity to process information, and in attempting to cope with this limitation individuals use heuristics or "short cuts" which usually lead to optimal solutions to problems but sometimes yield predictable errors (Abelson, 1976).

The impact of these sorts of experiment has been very great since the start of this era, perhaps started by the publication of Tversky & Kehnemann's paper in Science in 1974. The application of the knowledge of these shortcomings in inference to attribution theory (eg. Ross, 1977) has been particularly influential. Not only have several books concentrated on this theme (eg. Nisbett & Ross, 1980 and Kahnemann, Slovic & Tversky, 1982), but the influence has also spread throughout psychology in applied as well as pure research (see the discussion of the Hamilton (1980b) book about stereotyping, or Kinder & Weiss, 1978, for applications to foreign policy decision making, for example).

Snyder & Swann's paradigm fits into this popular framework too. Individuals are given the task of testing a hypothesis about another person, but their

estimates of those other people are influenced not only by the other person but also by the initial hypothesis they are given by the experimenter to test. Three important features of this bias in reasoning are common to the other work on shortcomings in human judgement. First, the observed errors are not random, but highly predictable. Second, features of the underlying epistemological processes can be inferred from the observed phenomenon. Third, the identification of these potential biases can be used directly to improve decision-making by taking steps to avoid the pitfalls once they have been identified.

Taken together, it can be seen that this programme of research was conducted *"in the right place at the right time"*. Its influence can be attributed as much to the research milieu in which it arose as to its intrinsic worth.

Summary

1. Snyder's programme of research into hypothesis testing about other people in social interaction is described. The process is charted from the way in which an "interviewer" prepares to test a hypothesis by attempting to elicit mainly confirmatory evidence, to the way in which "targets" oblige by giving confirmatory evidence, leading to the confirmation of the hypothesis.

2. The self-confirming hypothesis was then subject to many attempts to reduce or remove the effect. It was concluded by Snyder (1981) that it was a very robust phenomenon.

3. Snyder also describes how the propensity of individuals to find confirmation of their hypotheses can also be used to explain other phenomena such as belief perseverance and the perpetuation of social stereotypes.

4. Snyder's research has become very influential in Social Psychology. It is argued that this was at least in part due to the climate of research in the late 1970's. The assumptions, methods and conclusions of the research are compatable with the models of man, the treatment of emotion and motivation and the strong cognitive influence prevalent at the time. In addition Snyder's research complemented other theoretical research on the rationality of human inference (including attribution research) and other applied research on social stereotypes, both very popular topics for investigation.

In the next chapter some criticisms of the research into hypothesis testing will be put forward, which will form the basis of the experiment reported in chapter four.

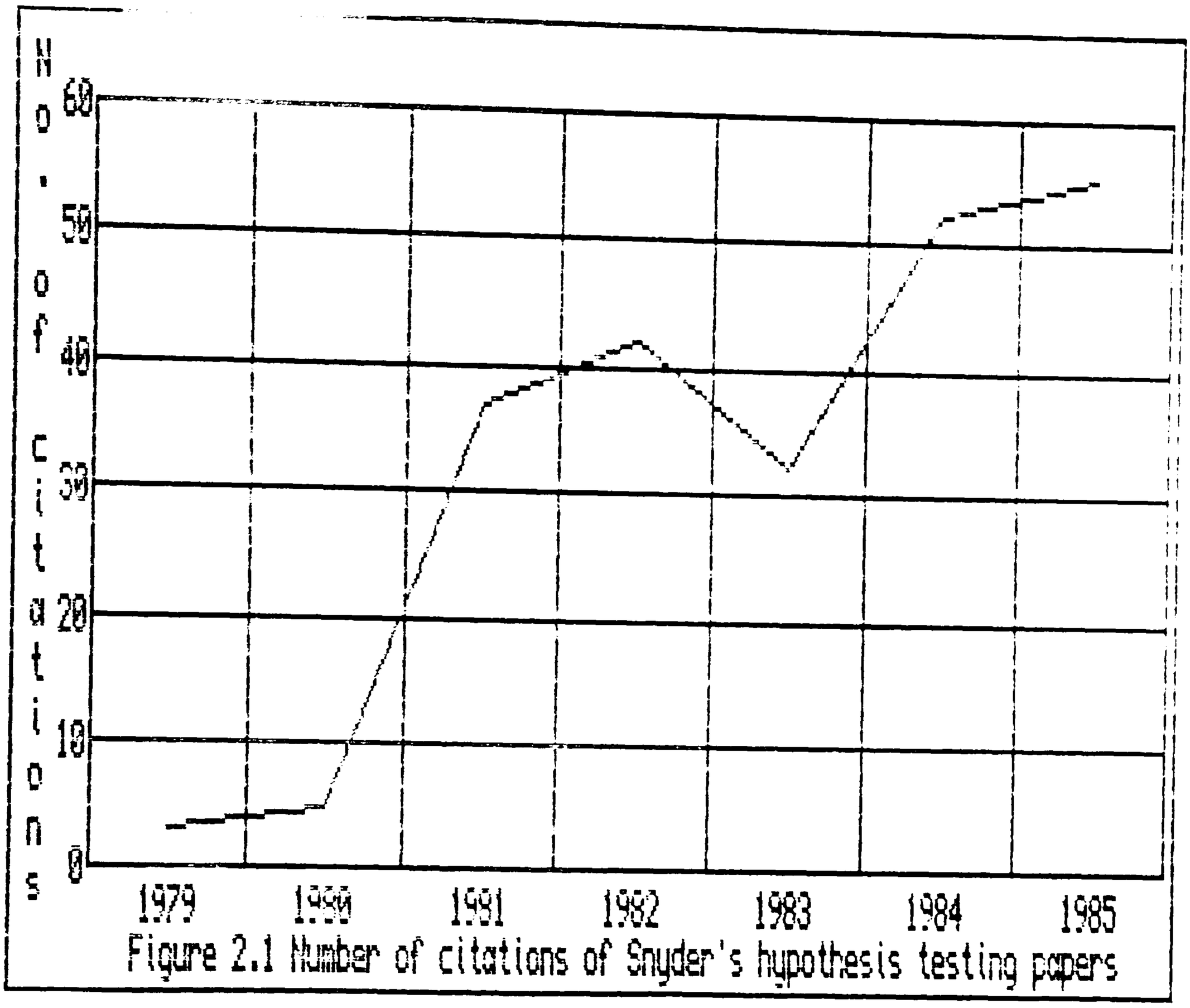

Figure 2.1 Number of citations of Snyder's hypothesis testing papers

Chapter 3

The Confirmatory Bias - From the perceiver's Perspective?

Introduction

In the last chapter Snyder & Swann's crucial (1978) experiment was discussed at length. One surprising omission in their experiment was that the interviewers, the people who actually asked the questions in order to form an impression of the "target", were never asked about the impression they formed of the target. The justification given for this was that other experiments had conclusively proved that hypothesis testers tend to see their hypotheses as having been confirmed. To quote Snyder & Swann's exact words...

"... But, did the interviewer-participants regard the hypotheses as having been confirmed by the actions of the target-participants? Although this investigation does not answer this question directly, other research (eg Swann, 1978) has demonstrated that after interacting with other people for the purpose of testing hypotheses, individuals do regard their hypotheses as having been confirmed."*

(1978, pp1207-1208)

*The work referred to here is Swann's unpublished PhD thesis.

This point about the interviewer's own impression of the target is again discussed by Snyder in a review of his work in 1981:-

> *"But, did the hypothesis-testers regard their hypotheses as having been confirmed by the target's actions? Apparently so. For, when all was said and done, the experimenter (during the post-experimental debriefing session) asked hypothesis testers what they had learned about their target's characteristic nature. Those who had tested the hypothesis that their targets were extraverts, on the average, regarded their targets as more extraverted by nature than did their counterparts who had tested the hypothesis that their targets were introverts."*
>
> (1981, p292)

These two quotes, both referring to exactly the same experiment, are clearly contradictory. The first statement infers that no measure of the interviewer's impression was taken while the second statement concerns such a measure, albeit, perhaps, an informal one.

The issue of bias and accuracy in the interviewer's final impression of the target is fundamental -- probably the most important issue in all of Snyder's experiments, which makes it all the more surprising that it is treated in such a casual manner. While the choice of confirmatory questions has been replicated many times

by Snyder & Swann (1978) in all four of their investigations as well as by Snyder and other collegues in follow up experiments (Snyder & Campbell, 1980; Snyder & White, 1981) and by other independent researchers (Cooper, 1980; Semin & Strack, 1980) there is no other published research that actually let the interviewers go on to ask their questions, or that measured the impressions that the interviewers formed of the target.

So, in order to look in more detail at the impression formed by the interviewers after interacting with the target and asking their questions, Swann's PhD thesis will be investigated in detail.

Swann's PhD Thesis Described

Swann's PhD thesis set out to look at the way in which individuals use social interactions to find out about other people, but the methods used are completely different from those used in Snyder's series of experiments investigating hypothesis testing processes in social interaction. Swann's experiment (the only empirical work in his PhD thesis) will be described before it is argued that, whatever conclusions can be legitimately drawn from Swann's experiment they certainly do not support the claim that interviewers are prey to a "confirmatory bias" when they test hypotheses.

Swann used males as the "*hypothesis testers*" and females as the *targets*. The males tested the hypothesis that a female they received prior information about and then interacted with either liked them or, in the other condition, didn't like them. The exact experimental procedure went as follows. The male subjects were told that the experiment was about how people get acquainted, but first they were asked to fill out an "attitudes and values" questionnaire which was taken away by the experimenter, and was (allegedly) shown to the female subject in the experiment. The male was told that this would help to make the interaction more realistic, because people generally know something about each other before they engage each other in conversation. In addition the males were told that it is normal for us to have some idea whether a person we are about to interact with will like us, but we are often fairly unsure about the accuracy of this information. In order to simulate this uncertainty the females (again allegedly) filled in a form to say whether, from the attitudes and values that the male had expressed through the questionnaire, she thought that she would like the male or not. This form was then shuffled in with two other bogus forms before the male took one of the forms. Thus the male was explicitly given the impression that while the form <u>may</u> have been from the female he was to interact with, this was by no means certain. It was then the male's task to find out whether the female really was favourably or unfavourably disposed towards him.

In fact the feedback forms did not originate from the female subjects at all but formed the 'hypothesis' manipulation in the experiment. In the *favourable hypothesis* condition the males received a form that indicated that the interaction partner thought that the male was "*moderately likable*" and "*strongly would*" like to get to know him better. By contrast, in the *unfavourable hypothesis* condition the form indicated that the partner thought that the male was "*moderately disagreeable*" and "*strongly would not*" like to get to know him better. The male was then given instructions to use the forthcoming interaction to find out whether what was written on the form accurately reflected what the female actually thought about him bearing in mind that it was probable that the feedback form was not in fact the genuine one filled in by the female they were going to talk with.

There was also a control condition in which the male subjects received no feedback about the female's opinion of them, nor were they given a specific hypothesis to test during the conversation. They were simply told to use the conversation to get acquainted with the female.

There was another factor in the experiment, orthogonal to the hypothesis manipulation. Right at the start of the experiment the male subjects were given a series of rating scales to assess their "self-

conceptions". The subjects were divided into two groups on the basis of this test. Those with scores above the median formed the "high self-perceived sociability" group; those with scores below the median formed the "low self-perceived sociability" group. Thus within each of the three hypothesis groups (favourable, unfavourable and control) half of the males typically expected their social interactions to go well, and the other males who generally saw themselves as being less likely to make a good impression in social situations.

The subjects were given four minutes from the time they received the feedback to prepare for the conversation. The experimenter then initiated the conversation by asking the two subjects to introduce themselves from seperate rooms using a telephone system. They were allowed to speak for approximately nine minutes. The interaction was tape recorded using a stereophonic recorder so that the male and the female's voices were recorded on seperate channels.

Independent judges, naive to the purpose of the experiment or the experimental condition of any of the pairs, listened to the tape-recorded conversations and rated them on several different behavioural and affective measures. This gave an indication of the strategies used by the males in testing their hypotheses. Other measures were taken after the interaction to see whether the males saw their hypotheses as having been confirmed. These were the

impression that the males thought that the females had got of them (ie. whether they thought that the females had liked them), and the impressions that the females had actually formed of the males.

The crucial test of the main experimental hypothesis, that the males will consider their hypotheses as having been confirmed, was tested by a 3 (*hypotheses*) X 2 (*percieved sociability*) analysis of variance using the males' ratings of the females impression of them as the dependent variable.

As predicted both main effects were significant. The males who had received favourable feedback in the hypothesis manipulation ended up thinking that the females really did like them more than the males in the "unfavourable hypothesis" condition, with the control condition coming approximately half-way between the two hypothesis conditions. There was also a significant "perceived sociability" main effect, showing simply that the males who rated themselves as being more socially adept thought that they had left the females with a better impression of themselves than the males in the "low self-perceived sociability" condition. A significant interaction was neither predicted nor found.

In the search for the mediators of this effect the judges' ratings of the males during the interaction were also used as dependent variables in the 2 X 3 ANOVA. Of the eight affective measures taken (agreement, praise or

compliments, interest, approval, tension, disapproval, disagreement, and antagonism) many failed to attain inter-rater reliabilities above 0.3 and the others either produced no significant effects, or the significant effects that were found bore no direct relevence to the interpretation of the results.

One other measure was taken by the judges, the frequency with which the males used a "*react to traits*" strategy. This is a conversational technique, identified by Swann, that individuals may use to test hypotheses. It consists of the male asking the female what she would think of a hypothetical person who possessed certain traits. Examples of this would be the questions "*What do you think of athletic men?*" or "*What do you think of doctors?*".

While Swann states that "*The measure of hypothesis testing was the frequency with which (males) use the 'react to traits' strategy*" (p.23), he does not make any predictions concerning this dependent variable at all. One plausible prediction is that the males in the hypothesis testing conditions would use this strategy more than those in the control conditions (i.e. a "*hypothesis*" main effect).

In fact the 2 X 3 ANOVA using this variable reveals no significant main effects. An interaction that is difficult to interpret indicates that the males in the "*favourable hypothesis / high self-perceived*

sociability" and the "*no hypothesis / low self-perceived sociability*" cells used this strategy more than the males in the other four conditions.

It is also informative to see what impressions the females finally had of the males. Were the hypotheses tru ly self-confirming? Did the females who interacted with males testing the hypothesis that they would get on well form better impressions of those males than the males in the "Unfavourable Hypothesis" condition?

No evidence was found of this; the main effect for the hypothesis manipulation was not significant. However, a significant interaction between self-perceived sociability and hypothesis revealed that high self-perceived sociability subjects made better impressions of themselves when testing a hypothesis (either favourable or unfavourable) than when in the control, no-hypothesis condition, whereas low self-perceived sociability subjects were not effected by the hypothesis factor. This finding is again difficult to interpret and of no direct relevence to the "self-confirming hypothesis" question.

Swann's PhD Thesis Criticised

Having given a description of Swann's experiment, it will now be evaluated to show that Snyder & Swann's (1978) claim that the results demonstrate that hypothesis testers will tend to see their hypotheses as being confirmed after social interactions to test those hypotheses, is unjustified.

The first mistake in their logic is in claiming that the males in this experiment were simply given hypotheses to test. In fact, they were given both a hypothesis to test and an expectation about the female's liking for them.

As has already been argued in chapter two, there is a crucial difference between hypotheses and expectations, and it is of fundamental importance in judging the rationality, accuracy and alleged biases of subjects testing hypotheses. A hypothesis is merely a testable statement about some feature of the world. It contains no information about the likely truth or accuracy of the statement -- it is merely a proposition. Thus, in their 1978 experiment, in most of the manipulations, Snyder & Swann gave the subjects no reason to believe that their hypotheses were any more likely to be true than false, they simply told the interviewers that "*the personality profile is a description of a type of person known to us all -- the extravert / introvert. You are to find out*

how well this profile describes the person you interview." (p 1204).

This is not the case in Swann's PhD thesis though. In this case subjects were not only told to "*test the hypothesis that this is how your partner feels about you*" (p 13) but were also given some information (the feedback from the "attitudes and values" questionnaire). Therefore the subjects not only had a hypothesis to test, but they were also led to expect that there was at least some chance that this hypothesis was more likely to be true than false. That is to say, it would have been normatively correct of the subjects in the "favourable feedback" condition to expect their chance of having a partner who found them likable to be higher than for the subjects in the "unfavourable feedback" condition. As in the Snyder, Tanke & Berscheid experiment (1977) the crucial test would be to see whether the males' representations of the females had diverged rather than converged after the interaction. Again measures that would allow this crucial test to be made were not taken.

This leads on to the possibility that when the males were asked how much the females had liked them after the conversation, they did not only use the impressions gained during the interviews, but they also incorporated this information gained before the interaction in their final estimation of the female's attitude to themselves. This would have been a

perfectly rational way for the males to integrate the information available to them, in accord with normative models of decision making such as Bayes' Theorem.

It has been pointed out by Nisbett & Ross (1980) that it is a common mistake in the psychology literature to assume that when individuals use preconceptions (in this case the unreliable evidence from the feedback form) they are being misled, and arriving at incorrect conclusions. For instance Kelley's classic (1950) "warm-cold" experiment is often cited as an example of subjects using information wrongly. Kelley did not draw this conclusion himself, and it is clearly not warranted.

Swann seems to be vaguely aware that this may be a problem; what he refers to as the males "merely parroting the hypothesis manipulation, without having made any attempts to test their hypotheses" (p 30). He offers several reasons why this is not the case, but none of his arguments stand up to scrutiny. His first mistake is to assume that <u>either</u> the subjects merely parroted back the hypothesis manipulation <u>or</u> they used the interactions to form their impressions; he does not consider the possibility that they used <u>both</u> bits of information and combined them using either a Bayesian model (treating the feedback manipulation as the prior odds) or some sort of an additive or averaging model. Either of these could have been normatively appropriate.

His first argument to support his position that subjects were not "*merely parroting the hypothesis*" is that because the subjects did use the "react to traits" strategy, they were obviously trying to test their hypotheses. Apart from the fact that, as stated in the previous paragraph, the subjects may well have been both making a serious attempt to use the interview as well as the hypothesis manipulation, a close inspection of the table of means shows that very few of the males used this tactic at all; the overall mean is less than 0.3 per subject per interaction -- only 28 occurences of this trait were identified by either of two judges in the all of the 97 conversations! A careful working back from cell means reveals that in fact only 18 of the 97 subjects were seen to have used the strategy <u>at all</u> by <u>either</u> judge. If, as Swann suggests, the use of the "react to traits" strategy is an indication of hypothesis testing, <u>one would be led to the conclusion that most of the males were not testing hypotheses</u>! What makes this argument even less convincing is the fact that the strategy was used by more of the males in the control or "no-hypothesis" condition (26.5%) than by the males in the experimental condition (14.3%).

Swann never gives any theoretical support to his use of this "react to traits" variable. It is difficult to see how it adds to an understanding of hypothesis testing process at all and seems to have been added into the experiment as an afterthought. He also used an analysis of variance model to analyse this variable, but

the parametric assumptions of the test were wildly violated -- two of the cells having zero variance for instance! Although Swann seems to have been oblivious to these problems, a re-analysis of Swann's data using a non-parametric analysis of variance (a technique outlined in Wilson, 1960) again shows no main effect but only a highly significant (but uninterpretable) interaction.

A second argument that Swann puts forward to defend his position is that there was a significant correlation between the males' estimates of the females' liking for them and the females' actual liking for them ($r_{(93)}=0.23$, $p=0.013$, one-tailed). Again, this is entirely consistent with the males using an averaging or Bayesian strategy to combine the information from the hypothesis manipulation and the conversation.

Swann continues with this argument by pointing out that the correlation is higher in the hypothesis testing conditions ($r_{(60)}=0.25$, $p<0.025$, one-tailed) than in the control, no-hypothesis condition ($r_{(31)}=0.11$, ns.).

Not only did an analysis of this difference using Fisher's Z_r transformation (See Ferguson, 1981, for further details of this technique) show that this difference was highly non-significant ($Z=0.62$, $p=0.27$, one-tailed) but it is not clear why one should expect a higher correlation in the hypothesis testing rather than the control conditions anyway. It could equally well

have been argued that one would expect higher accuracy in the no-hypothesis conditions since they did not have any bogus feedback information to confound their inferences!

There is a further important difference between Swann's (1978) experiment and the type of effect that Snyder & Swann are trying to obtain in their 1978 investigations. Swann clearly finds that, whatever the hypothesis testers conclude about their hypotheses, this is entirely an *"in the eye of the beholder"* effect. That is to say, there was no objective evidence that the females had actually come to like the males in the favourable hypothesis condition more than the males in the unfavourable hypothesis condition. In fact, a close inspection of the means indicates that the males in the unfavourable hypothesis condition were actually considered to be slightly (but not significantly) more likable and friendly by the females whom they interacted with than the males who were in the favourable hypothesis condition! If this small difference in means does reflect a real effect then there is a very interesting process occurring. The males may be compensating for the view they believe that the females hold of them, causing a "suicidal prophecy" -- the males who thought that the females did not like them initially came to be more liked.

While this surprising reverse effect is not significant, Snyder & Swann (1978) go to great lengths

to point out that the hypothesis testing procedure "*caused the targets to provide actual behavioural confirmation of the participants' hypotheses*" (p.1202). The discussion in chapter one (ibid) suggested that the conceptual difference between objectively and subjectively perceived support is not as simple as Snyder has assumed in his experiments, there does seem to be an important difference between the hypotheses being tested for in Snyder & Swann's experiment and the hypotheses being tested in Swann's thesis.

In testing for introversion or extraversion interviewers are testing for an enduring and already formed quality of an individual. By contrast, the liking that is being tested for in Swann's thesis is not a pre-formed trait but an affective and cognitive representation that is formed during the interview. While it is hardly likely that the interviewer could have a real effect on the target's enduring extraversion during a short interview, one's own behaviour clearly can have an important effect on other people's liking for oneself. Thus, any inferences drawn from Swann's thesis (even if it did find unequivocal evidence of a self-confirming hypothesis) and Snyder's introversion-extraversion paradigm must be carefully considered in the light of this difference. What is clear is that the uncritical way in which Snyder & Swann (1978) make inferences between one experiment and the next is unjustified and misleading.

To conclude these criticisms, Swann's (1978) data are entirely consistent with the notion that the males integrated all of the information available to them about the female's impression of them in a normatively appropriate and rational manner, and there is no evidence of a "confirmatory bias" on the part of the hypothesis testers at all.

To relate this back to the assertion by Snyder & Swann (1978) that hypothesis testers themselves are likely to consider the hypotheses they test as being confirmed -- this is clearly not proven, as Snyder claims it is, by Swann's (1978) work. Whether hypothesis testers are biased by the hypotheses in making judgments about the validity of their hypotheses remains an unanswered empirical question.

Summary and Conclusions

Snyder (1981, Snyder & Swann, 1978) does not provide any empirical support for the assertion that the interviewers in his experiments considered their hypotheses as having been confirmed. Instead he refers back to Swann's unpublished PhD thesis as proof of this.

Swann's PhD thesis was described in detail, then criticised for the following reasons:

1/ The concepts of hypothesis testing and expectancy utilisation which are crucial in understanding the tasks of the subjects in these experiments were confused. This meant that even the support that Swann claims to have found for self-fulfilling hypotheses is highly equivocal.

2/ There was, in fact, no main effect for the favourable / unfavourable hypothesis manipulation on crucial dependent variables -- the only significant effects were interactions that proved difficult to interpret.

3/ The measures taken to analyse the social interactions in the experiment (in particular the "react to traits" measure) were without construct validity, and conceptually confused.

4/ Many of the statistical techniques utilised by Swann to bolster his interpretation of the empirical findings are either wrongly applied or incorrectly interpreted.

5/ There were conceptual differences in the nature of the hypotheses used in Swann's PhD thesis and in Snyder's experiments (eg. Snyder & Swann, 1978) that should cast doubts on any direct generalisations from one of these experiments to the other.

Chapter 4
Hypothesis testing as a dynamic rather than static process.

Introduction

From the preceeding three chapters it can be seen that Snyder and his collegues have conducted a detailed series of experiments that have explored the tendency for people to test hypotheses by searching preferentially for confirmatory evidence. Several pieces of empirical work will now be presented that will continue the investigation into this phenomenon.

This first experiment is an attempt at a constructive replication of Snyder & Swann's second investigation in their initial, 1978, paper on the topic of hypothesis testing. Several key features of the experiment will be modified either to extend the understanding of the results or to correct inadequacies that were identified in the paradigm.

Following from the case made in the last chapter, the first modification was straightforward addition to the experiment. After the interviews have taken place, the interviewers were asked whether they thought that their hypotheses have been confirmed. This was done by getting them to use the same 10 bipolar scales used by the "naive rater-judges" in Snyder & Swann's experiment.

They also completed another measure of their estimation of the target's extraversion by filling in a version of the Eysenck Personality Inventory (EPI) as they thought that the target would answer the statements. With these measures it will be possible to determine whether the hypothesis-confirming phenomenon is a true "*participant experienced*" self-confirming hypothesis, or the less important "*observer experienced*" self-confirming hypothesis. The main reason behind using an additional second measurment is to aid in the estimation of the magnitude of the hypothesis effect. This will be discussed in more detail later in this section.

Another modification to the procedure involves the task given to the rater-judges. It will be recalled that in Snyder & Swann's 1978 experiment the rater-judges listened to only the targets' responses to the questions and not the questions themselves. The reason for this, though not made explicit in the report, was presumably to completely eliminate any possibility of the interviewer communicating the hypothesis directly to the rater-judges. Thus Snyder & Swann hoped to show that if the judges rated the targets as being more extravert in the extravert-hypothesis condition than in the introvert-hypothesis condition this can only have been because the interviewers made them behave in that way.

There is a fundamental flaw in the logic of this methodology. In looking at just the response without

the question that elicited that response, one is looking at behaviour in a vacuum without considering fully the possible causes of that behaviour. In the simplest analysis of the causes of a particular behaviour two possibilities have to be considered; either the behaviour was caused by something internal to the actor, or the actor was made to behave in that way by situational factors. This internal / external distinction is generally considered to be one of the most fundamental attributions that individuals make when analysing the actions of others (Kruglanski, 1983).

What Snyder & Swann have done by not giving the rater-judges access to the situational forces acting on the target is to make it virtually impossible for the judges to correctly attribute the targets' responses to either the dispositions of the targets themselves or to other, situational factors.

An example will help to make this point clear. Imagine that as a rater-judge one hears a target respond *"When I was 12 years old I went to the local youth club but I didn't get on with anyone there so I didn't go again. That was the only time that I can recall."* As the statement stands one might conclude that this reveals little about the personality of the target; one's isolated experiences at the age of 12 are very poor predictors of sociability as an adult. However, if this response came in answer to the question *"Tell me about any times that you joined a group or society. How*

well do you typically get on in those types of situations?" then the fact that the only example that the target could give of joining a group was when he was twelve years old, and then it was a failure, is very strong evidence of introversion. Conversely, if the question was *"Tell me about all the times that you have felt that you didn't get on well with a group"* then the fact that the target had to delve back so far into the past to find one, and only one, example of social isolation would probably lead one to the conclusion that she was usually extremely confident and proficient in social situations.

It can be seen from the theoretical analysis and from the hypothetical example that the practice of letting the rater-judges listen only to the target's responses is highly artificial and not representative of the usual information available to participants in conversations. Rather than making the perceptions of the targets by the rater-judges more objective, it has succeeded only in making them largely irrelevant. How often, after all, does one hear only one side of a conversation, the answers but not the questions?

This is exactly analogous to the argument put forward in Chapter 1, where it was pointed out that several expectancy-confirmation experiments also hid the perceivers' behaviour from the observers, again to the detriment of the generalisability of the experimental findings.

This is not to say that there are not also problems in allowing the rater-judges to listen to the interviewer's side of the conversation too. For instance Swann, Giuliano & Wegner (1982) found that rater-judges will form impressions of targets even when listening to just the questions asked of them by interviewers. Merely detecting the fact that the interviewer is probing for either introversion or extraversion can lead the rater-judges to infer that the interviewer is searching for that information with some justification.

Another consideration is to what extent will the rater-judges actually use the questions asked by the interviewer in interpreting the targets' responses? The evidence from the social cognition literature shows that individuals tend to ignore or under-utilise any situational factors and attribute the behaviour of others to their stable dispositions even when the situational forces are blatantly obvious. For example, Ross, Amabile & Steinmetz (1977) found that even when students were randomly assigned to the roles of questioner and contestant in a mock-up of a quiz show and the questioner was allowed to ask any questions he chose, the questioner was consistently rated as more intelligent than the contestant by both contestants and observers. The persistance and power of this effect whereby other people's behaviour is attributed to dispositional rather than situational factors has le d

to it being called the "Fundamental Attribution Error" (Ross, 1977).

A further addition to the measures taken will aid in the interpretation of the results, by analysing the amount of variance accounted for by the confirmatory bias. As was discussed in chapter 2 (ibid) Snyder has made some provocative inferences from his laboratory experiments to the real world, with quotes like "... *individuals may live in social worlds in which hypotheses become self-confirming hypotheses and beliefs become self-perpetuating beliefs -- social worlds in which beliefs can and do create reality.*" (Snyder & Gangestad, 1981).

In order to consider that an effect has importance outside of the laboratory one has to show not only that the effects observed are statistically significant, but also that they are of sufficient magnitude to make their contribution important (see, for example, Keppel, 1973 or Winer, 1971). While Snyder & Swann did not compute the variance accounted for by the hypothesis manipulation, a crude "eyeball test" of their results suggests that the effects are not large. For instance, only four of their ten measures associated with extraversion (introverted - extraverted, confident - unconfident, awkward - poised and energetic - relaxed) showed significant differences between the two conditions and the overall summated $\underline{F}$ ratio was only just significant ($f(1,38)=4.56$, $p=0.04$)). It could

well be that a weak effect could fade into irrelevance when compared with other sources of variance such as individual differences in the "acutal" extraversion of the targets or error on the part of the interviewers' perceptions. For this reason the subjects in the role of the target were selected to form two groups, an extravert group and an introvert group. Both groups had scores on the EPI (Eysenck Personality Inventory) at least one standard deviation from the mean. Thus a comparison could be made between the relative strength of the confirmatory bias phenomenon compared to the stable dispositions of the targets as measured by a proven personality assessment questionnaire. In addition, the interviewers were asked to fill in the questionnaire after the interview as they thought the target would fill it in. This formed another dependent measure, but one that was more helpful for the estimation of the magnitude of the effects.

Another simple methodological change was to allow the interviewer and target to sit face to face, rather than restricting communication to an audio-only link. It is not clear why Snyder & Swann chose to conduct the experiment in this rather artificial way; they do not give any justifications in their report, but the reasons were probably more historical rather than logical. In Snyder's two previous papers there had been good reason to separate the subjects during the interaction. The subjects in Snyder, Tanke & Berscheid (1977) were deceived about each other's physical attractiveness, so

face-to-face interaction was clearly not permitted to guard the experimental manipulation. Snyder & Swann's (1978b) subjects were separated because the experiment took the form of a prisoner's dilemma game, and the communication between the two "contestants" was limited to the bare minimum. Swann's PhD thesis (submitted in 1978 also) was also conducted over an audio-only link. The only justification given by either author for the use of this limited channelling of communication in these experiments (that are so atypical of most social interactions, telephone conversations being the one notable exception) is given in a later paper by Swann, Giuliano & Wegner (1982) where they say that it would "*prevent (the subjects) from employing nonverbal signals that would not be recorded on the audiotape.*" (p. 1028).

This being the case, why not use face to face interaction and videotape the proceedings for the judges? It may return some of the normality to an already artificial situation in the laboratory. There are, after all, well researched differences between interactions conducted over different channels of communication such as telephone conversations, face to face conversations and video-linked conversations (Short, Williams & Christie, 1976; Williams, 1977).

Intuitively one would expect that person-perception would be more accurate in face-to-face interactions than audio-only interactions because of the greater number of

cues available, particularly the very revealing non-verbal cues which one has less conscious control over. Who, for instance, would interview job applicants or interrogate suspects over the telephone?

Surprisingly, the consensus of research on the topic has found exactly the opposite. Inferences about the personality of others are more accurate the further the judge is removed from the non-verbal components of the message. In one experiment reported by Short et al, for instance, it was found that subjects were best able to discriminate between true and false statements by witnesses when just given a transcript of their evidence, and audio-visual contact was less accurate than just listening to a recording. The reason seems to be that judges tend to devote too much of their time to trying to utilise the non-verbal "leakage" and thus ignore the more valid verbal and factual components of the message. Whether this would still be the case in this experiment where there in no motivation to deceive on the part of the target is less clear. Most studies that have explored this issue have found no significant effects of medium of communication on accuracy, although liking is often found to increase with the richness of non-verbal cues. Again, it is difficult to predict what effect, if any, this manipulation will have on a hypothesis-confirming bias.

One final change was made to Snyder & Swann's methodology, again to make the experimental situation

more representative of the kinds of situation in which people do try to find out about each other. In all of the previous experiments on hypothesis testing the subjects chose all of their questions to ask <u>before</u> they even met the target, so their part in the interaction consisted of simply presenting the pre-determined questions in a pre-determined order. This is very atypical of most social interaction. Even in very formal semi-structured interviews only some of the questions that the interviewer will ask would have been prepared beforehand. The rest would evolve during the conversation, depending on the target's answers to previous questions. In less formal conversations where people attempt to form impressions of each other it is doubtful whether the participants would have anything more than a vague idea of some of the topics of conversation they may evoke. It has often been argued that one of the major criticisms of Social Psychology is its attempt to reduce complex, dynamic processes and analyse them as if they were simple decisions. In this situation such artificiality is not necessary, the interviewers can choose which questions to ask as the interaction proceeds. It can then be seen whether the interviewers' questioning strategies are affected by the cognitive pressures of simultaneous speech, and whether their strategies change over the course of the interaction as they develop an increasingly complex representation of the target's personality. It is predicted that this modification will lead to a reduction of the confirmatory bias over the course of

the interview. Whereas the interviewers are likely to start off by asking confirmatory questions in the first half of the interview, as they become more involved in the situation the power of the confirmatory bias will be reduced. This will show itself by an interaction between the hypothesis under test and the order of the questions in the interview on the types of question asked. Other variables, such as the personality of the target may also interact with the ordering of questions in the interview such that extravert targets come to be asked more extravert type questions in the second half of the interview (and vice-versa for introvert targets) as the interviewer comes to concentrate more on the characteristics of the target than on the nature of the hypothesis.

The net effect of all of these changes (and a few very minor ones outlined in the method section) is difficult to predict. On balance though there are probably as many changes that may lead to a bolstering of the effect as ones that may reduce the effect. The main reasons for running this first experiment is thus to replicate the confirmatory bias in social interaction (given that only one of the many experiments on question selection for hypothesis testing have actually allowed the subjects to continue to ask the questions), to check Snyder's claim that the hypothesis-confirmation is experienced by the interviewer as well as by the rater-judges and to see whether the effect persists under more naturalistic conditions.

Method

Subjects

Forty male and 40 female undergraduates from the University of Birmingham volunteered to participate in the experiment. Half of these had already been selected for the role of target by virtue of their extraversion scores being one standard deviation or more from the mean for the population.

A further 27 subjects were selected from the University of Warwick to act as rater-judges.

Procedure

The questions used in this study were selected from the list of 26 questions but were re-classified into three groups using 14 judges (Snyder & Swann used only nine). The three groups were:-

Questions one would ask of someone already known to be an extravert.

Questions one would ask of someone already known to be an introvert.

and a neutral category for the remainder.

Snyder & Swann used a simple majority (ie. five or more judges) to assign a question to one of the two former categories, otherwise they were assigned to the last category.

The criteria used here were more stringent; a two-thirds majority was required for inclusion into the first two categories. The neutral questions were then excluded since there tended to be low-consensus about these questions, and they were not of theoretical interest. This left a list of 18 questions, nine extravert and nine introvert.

The subjects in the role of interviewer were given exactly the same instructions as in Snyder & Swann's second investigation. More specifically, they were given a card with either the personality profile of an introvert or the personality profile of an extravert on it, and were told that it was their task to find out whether the person that they were to meet was of the type described on the card. They were given no reason to suggest that this was likely to be an accurate description of the target, as in Snyder & Swann's "low certainty" condition. They were then told that they would find out about the target by asking questions of the target, and were given the list of 18 questions on a sheet entitled "Topic areas often covered by interviewers" (a copy of these questions can be seen in Appendix 4.1). They were told to study this list carefully, but to only choose one question to ask first

-- they would decide which other questions to ask as the interview proceeded.

The targets were simply told that the other person would be asking questions in order to find out about them, and they were instructed to answer them in as open and informative way as possible. To simplify things and eliminate further sources of variance the targets were all matched with an interviewer of the same sex.

The two subjects were then led into the same room, introduced and a check was made to ensure that they did not know each other except by sight. The interaction was then conducted over a round table in the laboratory, and was filmed with the use of two video cameras, one opposite each subject. These channels were integrated using a video-editor, to give one recording of a split-screen picture with one subject in each half, both facing their respective cameras. Although they were informed that they would be filmed the cameras were made to be as unobtrusive as possible.

After the interaction the subjects were led into seperate rooms again, and the interviewer-subject was asked to fill in the ten item bipolar scales, identical to those used by Snyder & Swann. The ten scales were:-

Talkative - Quiet

Unsociable - Sociable

Friendly - Unfriendly

Poised - Awkward

Introverted - Extraverted

Enthusiastic - Apathetic

Shy - Outgoing

Energetic - Relaxed

Cold - Warm

and

Confident - Unconfident

As well as being ten independent measures Snyder & Swann found that these scales had a high internal reliability (Coefficient alpha = 0.95) so could be summed to obtain one measure of general extraversion. One minor change was made in the administration of the test; instead of being six-point scales they were changed to 80 millimetre lines which the subject marked at an appropriate point to signify his or her estimate of the target's personality on that dimension. The form was scored using a key that divided the lines into 18 sections. Summed over the ten scales this gave scores on a 170 point scale from -85 to +85, higher scores indicating a greater degree of extraversion.

Then the interviewers were given a copy of the EPI to fill in for the target, being instructed to answer the questions as they thought that the target would answer them using the knowledge that they had gained from talking with them. Twenty-four of the 57 questions related to the scale of extraversion, the questions requiring simple yes/no answers. Examples of the

questions are "*Would you do anything for a dare?*" and "*Do you like talking to people so much that you never miss a chance of talking to a stranger?*". The range of possible scores on this test is from 0 to 24, higher scores being nearer the extraverted end of the range.

The subjects were then individually debriefed and thanked for taking part.

This part of the experiment was a fully randomised 2 (personality of target) X 2 (introvert hypothesis vs extravert hypothesis) X 2 (sex of dyad) design with five dyads per cell.

The next stage in the experiment was to see whether rater-judges who saw the interactions but were naive to the hypothesis being tested by the interviewer also saw the actions of the targets as being evidence in favour of the hypothesis.

Twenty-seven rater-judges were used, all undergraduates, postgraduates and staff of the University of Warwick. This ensured that none of the rater-judges were familiar with any of the interviewers or participants. Each rater-judge saw videotapes of ten interviews shown in random order, with five from the extravert target and five from the introvert target condition, but all within one hypothesis condition. Also, all rater-judges saw only dyads of the same sex as themselves. After seeing the recording of each

interview they rated the target using the same two measures as the interviewers, the EPI and the ten bipolar scales.

In order to ensure that the judges were diligent in their tasks a lie scale was included in with the extraversion questions. Rater-judges who scored over seven on the nine point scale were eliminated from the analysis.[1] Seven rater-judges were eliminated using this criterion, leaving 20. With 40 dyads in total, and 20 rater-judges seeing ten interactions each, each interaction was seen by five rater-judges.

The time taken by the judges to see and rate all ten interactions was about two and a half to three hours. This stage of the experiment was completely automated so a computer gave all of the instructions, administered all of the questions and told the rater-judges when to fill in the bipolar scales and go on to the next video-recording. This reduced any chance of experimenter effects and allowed the subjects to complete the tasks in their own time without feeling hurried.

1 For the first 14 subjects to run, the experiment was automatically terminated by the computer program as soon as a score on the lie scale exceeded six on the lie scale. The last three subjects who exceeded the limit of more than six positives on the lie scale did complete their allotted 10 interviews, and when their data, along with the data from whatever interviews were rated by the other four "rejected judges *were* analysed *they* showed the same overall pattern as the 20 other judges.

Results

For the sake of clarity the results section will be divided into three sections:- the questions asked: the impressions formed by the interviewers: and the impressions formed by the judges. As no sex effects were predicted or found in any of the results the sex factor will be ignored in the descriptions of the data and analyses (with one minor exception that will be discussed later).

The Questions Asked

An analysis of the number of extravert questions asked in the various conditions found no significant effect for the hypothesis factor ($F(1,32)<1$, n.s.), Nor was there a significant effect for the personality of the target, $F(1,32)=1.57$, n.s.). (Table 4.1).

The analysis for the introvert questions is an exact replication of this since the number of introvert questions selected was necessarily nine minus the number of extravert questions asked, there being no neutral questions.

However, a more detailed analysis of the results revealed a very interesting trend. The data was re-analysed incorporating the order in which questions were asked as a fourth, repeated measures factor in the analysis of variance by separating the first four from

the final five questions asked. There was now a significant interaction between hypothesis and order on the proportion of extravert questions asked, $F(1,32)=3.57$, $p=0.03$, 1 tailed). This indicates that the subjects started by asking predominantly questions that probed for confirmatory evidence in both conditions, but then this trend was reversed. The number of confirmatory questions asked thus went from 59% down to 46% (see figure 4.1). The difference in the proportion of extravert questions asked in the first four between the two hypothesis conditions was not, however, significant ($t(38)=1.10$, n.s., 1 tailed).

The interviewers' impressions of the targets

After the interviews the interviewers rated the extraversion of the targets by filling in the ten item bipolar scales and the EPI for them. The bipolar scales were found to have a very high internal consistency (coefficient alpha = 0.91), and were thus summated to form one overall measure. The results for the summated scales and the EPI are shown in tables 4.2 and 4.3 respectively. There was a highly significant main effect for the personality of the target on both measures, the extravert personality group being perceived as more extravert than the introvert personality group (Summated Scales: $F(1,32)=14.4$, $p=0.001$; EPI: $F(1,32)=28.9$, $p<0.001$).

There was, however, no main effect for the hypothesis factor (both *F*s<1), and none of the interactions approached significance.

The proportion of the total variance accounted for by the various effects was computed using the R^2 statistic described by Vaughan & Corballis (1969). The personality of the targets accounted for 41% of the total variance using the EPI or 25% of the variance with the summated scales. For both of the measures the mean square of the hypothesis factor was less than the mean square of the error term (i.e. the *F* ratios were less than 1) indicating that the proportion of the total variance accounted for by this factor was exceedingly small or non-existent.

One more analysis was attempted to detect any hypothesis manipulation effect. By removing the variance attributable to the exact self-administered EPI scores of the targets the error term would be reduced giving a much more powerful test more likely to detect even the weakest of effects. Even when this was done with an analysis of covariance (using the target's self-administered EPI score as the covariate) the hypothesis factor was still not significant, although there was now a non-significant trend, $F(1,37)=2.36$, $p<0.1$, 1 tailed), with the small difference between the adjusted means in the predicted direction.

The rater-judges' impressions of the targets

The rater-judges' impressions of the targets were very similar to the interviewers' impressions. Again there was a strong main effect for the personality of the target indicating that the introverted targets were seen to behave in a more introverted manner than extraverted targets (Summated scales: $F(1,32)=12.8$, $p=0.001$; EPI: $F(1,32)=29.9$, $p<0.001$). The means can be seen in tables 4.4 and 4.5.

The rater-judges' perceptions of the targets were uneffected by the hypothesis being tested by the interviewers; both F ratios were again below 1.

The personality of the targets as measured by the self-administered EPI again accounted for a large proportion of the total variance in the rater-judges' impressions of the targets on both the EPI (42%) and the summated scales (23%). As before, since the F ratios for the hypothesis effect were less than 1 the proportion of the variances accounted for by the hypothesis manipulation was at most very small.

Again an analysis of covariance was performed on the data as a last attempt to detect the effect of the hypothesis manipulation on the rater-judges' impressions of the targets. It should come as no surprise by now that the hypothesis factor was still not significant,

but again there was a slight trend in the predicted direction, $F(1,37)=2.72$, $p<0.1$, 1 tailed.

One surprise finding was that the interaction between sex and hypothesis was significant with the summated scale measure ($F(1,32)=4.4$, $p=0.041$) but not the EPI($F<1$). Given that this effect is only just significant, and on only one measure, it is quite likely to be a *type one* error. Furthermore it is of no clear theoretical interest, nor does it help in understanding the results.

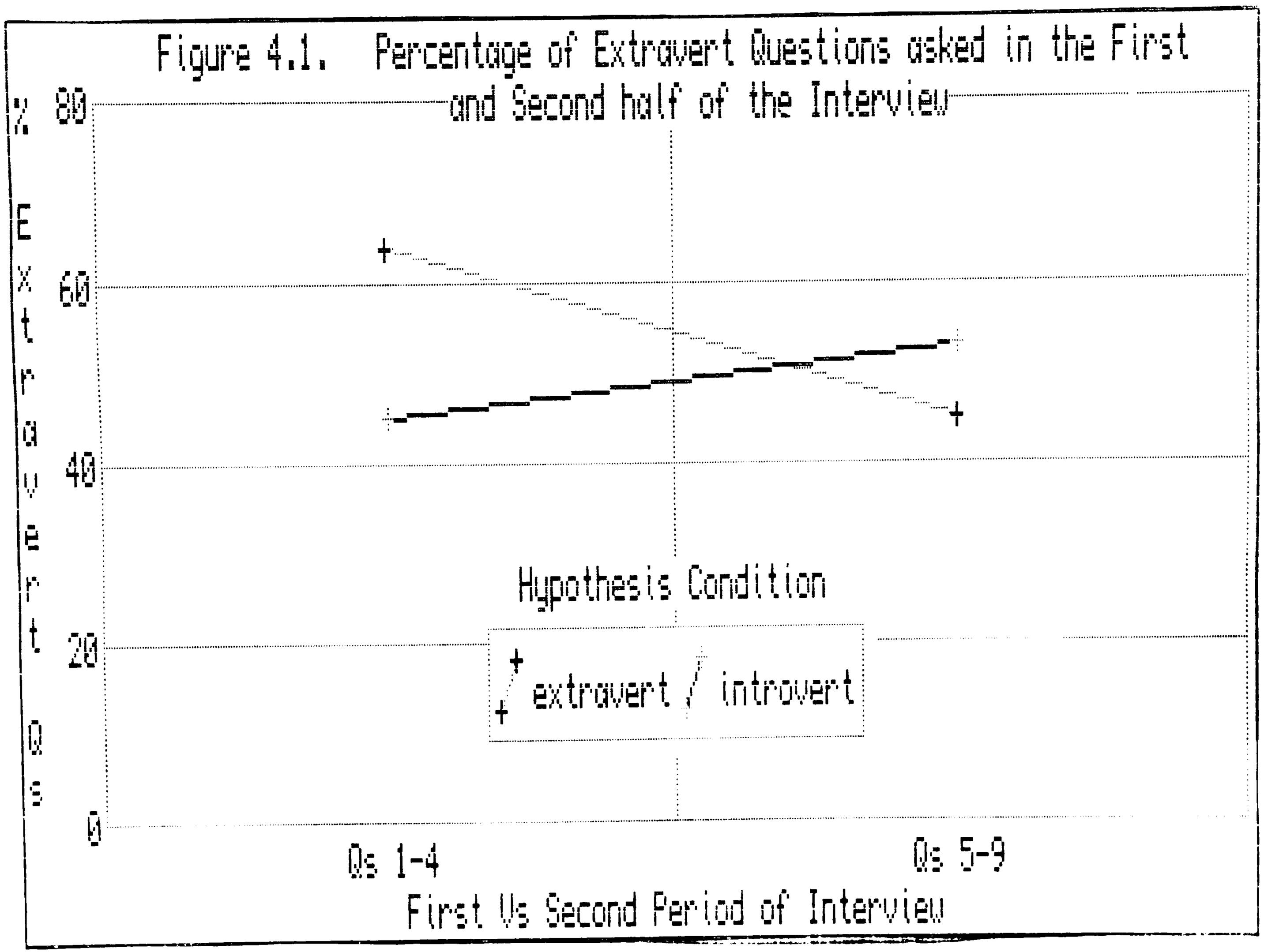
Figure 4.1. Percentage of Extravert Questions asked in the First and Second half of the Interview
% Extravert Qs
80
60
40
20
0
Hypothesis Condition
extravert
introvert
Qs 1-4
Qs 5-9
First Vs Second Period of Interview

Table 4.1. Mean Number of *Extravert* questions Selected

Hypothesis	Target's Personality: Extravert	Target's Personality: Introvert	
Extravert	M = 4.8 SD= 0.9 n = 10	M = 4.8 SD= 1.0 n = 10	M = 4.8
Introvert	M = 4.9 SD= 0.9 n = 10	M = 4.0 SD= 1.6 n = 10	M = 4.4
	M = 4.9	M = 4.5	

Table 4.2. Interviewers' Perceptions of the Target Using the Summated Scales

Hypothesis	Target's Personality: Extravert	Target's Personality: Introvert	
Extravert	M = 29.0 SD= 30.6 n = 10	M = 7.7 SD= 14.6 n = 10	M = 18.4
Introvert	M = 31.5 SD= 23.6 n = 10	M = -1.9 SD= 12.8 n = 10	M = 14.8
	M = 30.3	M = 2.9	

Table 4.3. Interviewers' impressions of the targets Using the *EPI*

Hypothesis	Target's Personality: Extravert	Target's Personality: Introvert	
Extravert	M = 14.8 SD= 6.4 n = 10	M = 5.7 SD= 4.9 n = 10	M = 10.2
Introvert	M = 13.7 SD= 6.2 n = 10	M = 3.6 SD= 3.7 n = 10	M = 8.7
	M = 14.3	M = 4.6	

Table 4.4. Rater-Judges perceptions of the Targets using the Summated Scales.

Hypothesis	Target's Personality: Extravert	Target's Personality: Introvert	
Extravert	M = 20.5 SD= 19.2 n = 10	M = 3.0 SD= 20.1 n = 10	M = 11.8
Introvert	M = 21.0 SD= 18.7 n = 10	M = -2.2 SD= 16.0 n = 10	M = 9.4
	M = 20.8	M = 0.2	

Table 4.5. Rater-Judges perceptions of the Targets using the *EPI*.

Hypothesis	Target's Personality: Extravert	Target's Personality: Introvert	
Extravert	M = 14.0 SD= 4.0 n = 10	M = 8.1 SD= 4.4 n = 10	M = 11.1
Introvert	M = 13.9 SD= 4.4 n = 10	M = 6.5 SD= 2.2 n = 10	M = 10.2
	M = 14.0	M = 7.3	

Discussion

These results are clearly at odds with those found by Snyder & Swann (1978). Not only did the interviewers not display a confirmatory bias in the questions they asked but, at the end of the day, they also failed to be misled into thinking that their hypotheses had been confirmed (or, *if* the non- significant trend from the analysis of covariance did turn out to be a reliable effect, the effect was so small as to be trivial).

The possible reasons for these differences will be discussed in two sections -- the questions asked and the impressions formed -- before the implications for both the concept and relevance of the "confirmatory bias" are considered.

The Questions People Ask.

Snyder and his collegues conducted a long and arduous series of experiments in the attempt to find any influences that would limit the power of the confirm-atory bias. Three years after his first published paper on hypothesis-testing Snyder concluded "*If any procedure exists for inducing individuals to eschew confirmatory hypothesis-testing strategies in favour of either disconfirmatory or "equal-opportunity" hypothesis -testing strategies, that procedure has yet to appear.*" (1981, p.290).

The next year he rejoiced in having found such a strategy, but that was only after having gone to extraordinary lengths with various educational and "*impression management*" interventions (Snyder, Campbell & Preston, 1982).

Yet in this experiment the interviewers formulated "equal-opportunity" strategies under very normal, everyday circumstances. There are two possible expla nations for this non-occurrence of the confirmatory strategy. Either it was because the list of questions was different from those used in Snyder's series of experiments, or it was something to do with selecting the questions *during* rather than *before* the interaction.

There are two reasons why the non-occurrence of the effect was probably not simply caused by the different questions. Firstly, other experimenters have also found the phenomenon using their own lists of questions developed and categorised independently from Snyder's list -- and a lot further removed from Snyder's list than the list used here (Cooper, 1982; Semin & Strack, 1980). It would thus seem unlikely that there were some particular artefact of Snyder's question list that caused the confirmatory bias. Secondly, there is evidence that the interviewers in this experiment started by asking more confirmatory questions but that it was something that occurred during the interaction

that made the interviewers reject their confirmatory hypothesis-testing strategy.

This leads on to the other possible expla nation -- that the hypothesis-confirming strategy has been interfered with by making the question choice occur simultaneously with the interview, rather than making the interviewer select all of the questions before the start of the interaction.

An analysis of the information available to the interviewers before and during their question selection task shows why this small change in methodology may have influenced the types of question chosen by the interviewers to test their hypotheses.

One of the main lessons to be learned from the social cognition literature is that individuals are often influenced by salient, vivid or concrete information at the expense of abstract, dull or pallid information (eg. Reyes, Thompson & Bower, 1980 -- See Nisbett & Ross, 1980, Ch. 3 or Fiske & Taylor, 1984, Ch. 7 for reviews of the literature on this point.). When the questioning strategy required interviewers to select questions before their interactions with the targets, all they had to go on were the experimental instructions and the personality profiles of a typical introvert or extravert. In such a situation these instructions would have been utilised to the full in selecting the questions. It is even likely that the interviewers used

a "matching" heuristic in deciding which questions to ask. This is a strategy identified by Evans (1972 and 1980) by which subjects, faced with a complex task, simply choose responses that have the same surface appearance as the phrasing of the task itself. In other words, the interviewers testing, say, the introvert hypothesis would identify the task as being "something to do with introversion" and then select out any questions on the theme of introversion.

However, when selecting questions *during* the interview the information available to the interviewer is both more complex and more plentiful. The target is quickly going to take over from the hypothesis as the most vivid feature of the environment; as Heider said "Behaviour ..... has such salient properties it tends to engulf the total field" (1958, p. 54). The crucial factor in determining which question to ask is rapidly going to shift from the hypothesis to the target.

While this explanation seems very plausible there is a minor point that detracts from it. It would be reasonable to expect that, if the main factor in the choice of questions had transferred to the target from the hypothesis, then there would either be a main effect for personality or an interaction between order and personality such that the interviewers came to ask more extravert type questions of extraverted targets and more introvert type questions of introverted targets. There was no evidence of this at all; in fact both of these

F-ratios were less than one. This may not go entirely against this theory, since other experimental evidence shows that individuals do not search for confirmation of their beliefs, only their hypotheses (Semin & Strack, 1980; Merteens, 1984). This distinction, often confused in the literature, was discussed in greater detail in chapter 2. One puzzle that remains is to determine what strategies interviewers are using when selecting questions if the types of questions chosen are unrelated to either the hypothesis under test or the personality of the target.

There are, in fact, several other strategies that the interviewers could be using in making their question choices. For instance they may be selecting what they perceive to be the most diagnostic questions to ask, a strategy that has been shown to be more prevalent and to account for more of the variance than the confirmatory strategy (Trope & Bassok, 1983). Another strategy for choosing questions that has been identified follows from considerations of social desirability. Cooper (1982) assumes that interviewers probably do not want to seem rude or make themselves or the targets feel uncomfortable by asking awkward questions. He asked rater-judges to rate each question on a seven point scale to indicate how comfortable or uncomfortable they would feel asking that question of a stranger. It was found that the Snyder & Swann's introvert questions were, on average, rated as being more uncomfortable to ask than the extravert questions. Cooper predicts that this may be

the reason why, over all of these experiments, the introvert questions are generally selected less often than the extravert questions.

One other possibility is that the lack of a confirmatory questioning strategy could have been an artefact associated with the modified and shorter list of questions. As explained earlier, the neutral questions were removed, the list was trimmed down to 18 from 26 questions and interviewers were told to select only nine (as opposed to 12) of the remaining questions.

Both the lack of a confirmatory bias and the order effect may also have been an artefact of the mathematics of selecting such a large subset of items, without replacement, from a finite list. For instance, an interviewer may have looked down the list and rejected all of the questions they considered to be embarrassing to ask or non-diagnostic (that is to say, of no use in differentiating between introverts and extraverts). This may only have left about half of the questions deemed to be acceptable, so there would have been, in effect, a "Hobson's choice" between questions searching for confirmatory or disconfirmatory evidence. Moreover, the interviewers may have "exhausted" the acceptable confirmatory questions first, then been forced to use up the disconfirmatory questions, thus accounting for the order effect.

A post-hoc test was performed to look into this possibility. It was reasoned that if this were the case, and if there were a high consensus between the interviewers as to the good and poor questions, there should be clear evidence that some of the questions were utilised reliably more often than others. A Chi-square test revealed that this was not necessarily the case ($X^2(17)=21.97$, $p>.2$), but this is only an indication. It may be the case that while the interviewers were using this "*eliminate then choose*" strategy, there was simply no consensus as to which questions were considered preferable. A true test will be to see whether interviewers choose confirmatory questions from the exact list used here, but when they have to choose before rather than during the interaction, that is to say under similar conditions to those in Snyder & Swann's (1978) experiment. This will be checked in the next experiment by using the same 18 questions as in this experiment, but with the question selection stage preceeding the interview.

If, however, it is found that the critical variable is the interaction during the question selection, then the important question now becomes which of the two methodologies is more representative of the hypothesis testing processes in people's unconstrained everyday environments?

As discussed earlier, situations in which the questioner would have a fully prepared set of questions

to ask before an interaction are exceedingly rare. Television and radio interviews by journalists may sometimes seem to take this form, and look how unrepresentative of normal intractions they seem! Have Snyder & Swann (1978) overlooked one essential feature of social life; that, rather than being discrete actions and decisions, it unfolds as a dynamic and ongoing process?

The results of the next experiment are necessary before this conclusion can be verified.

The interviewers' perception of the target.

Virtually no evidence was found to suggest that the interviewers' perceptions of the targets were biased by the hypotheses they were testing. This lack of an effect was unlikely to have been caused by by a lack of sensitivity of the measuring instrument or a lack of diligence on the part of the interviewers given the very strong effect for the personality of the target.

This finding is also contrary to those of Snyder & Swann (1978), although from the critique of the evidence presented by Snyder & Swann and Snyder (1981) in chapters 2 and 3 of this thesis, it is perhaps not surprising that the confirmatory bias was not found. There is an obvious reason why the interviewers in this experiment did not tend to conclude that their hypotheses had been confirmed. Snyder's hypothesised mechanism for the confirmation of hypotheses requires the interviewers to ask questions that seek out confirmation of the hypothesis. Because the method of question selection prevented there being a confirmatory bias in the questions asked in this experiment, this confirmatory mechanism was necessarily ruled out.

It is impossible, therefore, to know whether the interviewers would have thought that their hypotheses were confirmed if they had asked a majority of questions that sought out confirmatory evidence. One tempting way to answer this question would be to re-analyse the data

collected looking for differences between the interviewers who asked more extraverted questions and interviewers who asked more introverted questions. This would not, however, give an unequivocal answer since the reasons why interviewers chose the questions they did is not known, but there may well have been good reasons for different questioning strategies that were not independent of the target's personality or behaviour in the interview situation.

Unfortunately the question of whether the self-confirming hypothesis phenomenon caused by an imbalance in information search is a "*participant experienced*" as well as an "*observer experienced*" confirmation remains unanswered by the data collected from this experiment. A further experiment in which the interviewers are induced into asking confirmatory questions must be conducted in order to answer this question.

The rater-judges' perceptions of the target.

The rater-judges did not detect any differences between the targets in the introvert hypothesis condition and the targets in the extravert hypothesis condition. It would have been very surprising, in fact, if the rater-judges had found there to be any difference between the hypothesis conditions given that there was no difference between the types of questions asked in the two conditions, and that they were naive as to the

nature of the hypothesis under test anyway.

It is worth noting again though how the strong effect for the personality of the target accounted for a moderately large proportion of the variance. This demonstrates that the rater-judges were doing their task at least somewhat conscientiously and that the lack of the hypothesis effect could not have been caused by the insensitivity of the measures.

General Discussion

It seems as if the lack of any of the effects for the hypothesis manipulation predicted by Snyder & Swann (1978) can be attributed wholely to the way in which the interviewers chose their questions. It is interesting to note that the predisposition of subjects to seek confirmation of their hypotheses may not occur in social interaction at all, but only under very artificial laboratory conditions.

It is still of theoretical interest, however, to find out whether interviewers would consider their hypotheses as having been confirmed *if* they *did* ask a majority of questions that preferentially attempted to seek out evidence supportive of the hypothesis. This question will be addressed directly in the next experiment.

Conclusions

It was found that when the interviewers selected their questions to ask *during* the interaction with the target there was no overall confirmatory bias in the questions asked. Furthermore there was an interaction between the types of questions asked and the order of the questions; interviewers were more likely to ask confirmation-seeking questions in the first half of the interview than in the second half.

It was argued that this effect was probably due to the presence and salience of the target, over-riding the hypothesis manipulation as a factor in question evaluation and selection. This cannot be proved conclusively, though, until the confirmatory bias in question choice is replicated using the same list of questions as employed here, but with the selection task preceeding the interaction between target and interviewer.

The hypothesis under test did not affect the interviewers' or the rater-judges' perception of the target. However, as there was no difference in the questions asked between the two hypothesis conditions, it is not possible to generalise this finding to situations where a confirmatory evidence is sought. The next experiment will address this question too.

Chapter 5. A Second Attempt to replicate the Confirmatory Bias

Introduction.

The Questions Asked

The previous experiment found that the self-confirming hypothesis effect did not replicate when several changes were made to the experimental procedure. It was not clear exactly which of these changes was responsible for over-riding the effect, but the lack of a predominance of confirmation-seeking questions was likely to have been a principal factor in eliminating the effect.

While it seemed plausible that the absence of the confirmatory bias was caused by the interaction with the target during the choosing of the questions, there was another possible explanation that could not be ruled out; that it was an artefact of choosing a larger subset of questions from a finite list. In order to eliminate this possibility, the same set of questions will be used in this experiment, but the interviewers will select their questions to ask *before* the interaction with the target. If, under these conditions, the interviewers do select confirmatory questions then that will demonstrate that it must have been the interaction with the target that removed the confirmatory bias in questioning

strategy in the last experiment.

The Impressions Formed

If the confirmatory choice of questions is replicated, this experiment will allow further questions to be answered. The most important question is whether, if the interviewers seek out evidence in favour of their hypotheses by their choice of questions, they will be "fooled" into thinking that the data is representative of the targets' true dispositions rather than simply a function of the questions that were asked of them? This weak link in Snyder & Swann's confirmatory process was described in Chapter 3, and still has not been properly tested due to the lack of the confirmatory bias in the questions asked in the previous chapter.

This question can be answered simply enough by letting the interviewers continue to ask the questions that they selected, and then by determining whether the impressions the interviewers have formed of the targets were dependent upon the hypotheses that the interviewers were testing.

If the interviewers did perceive their hypotheses as having been confirmed, the audio and video recordings of the interactions could be analysed to shed more light on the exact mechanisms that cause the effect. As Darley & Fazio (1980) argued, two conceptually different mechanisms can be identified that could lead to self-

fulfilling prophecies occurring. The first occurs when individuals are presented with mixed or ambiguous evidence, and they interpret it as being consistent with the expectation. This may occur because individuals might preferentially recall confirmatory evidence or weight confirmatory evidence more strongly than disconfirmatory evidence. The second type of self-fulfilling prophecy occurs when individuals act in such a way as to collect an unrepresentative set of evidence which would appear to confirm the prophecy even to someone who was unaware of that expectation.

There are several examples of the first type of self-fulfilling prophecy in the literature. For instance, Rothbart, Evans and Fulero (1979) found that when subjects were given expectancies about the likely behaviours of a group of men, they would recall more of the expected types of behaviour and thus consider that their initial expectancy was indeed true. Similarly, Snyder & Uranowitz (1978) showed how preconceptions -- this time brought about by stereotyped expectations of homosexual as opposed to hetrosexual women -- could bias the recall and processing of mixed evidence. There is also empirical evidence that exactly the same phenomenon can occur in hypothesis-testing. Snyder & Cantor showed again the testing of a simple hypothesis could effect the recall of previously learned material (1979).

To see whether the effect Snyder & Swann claim to have discovered is truly an example of a situation where

the interviewers have <u>created</u> evidence that would convince individuals who saw just that evidence and were not aware of the hypothesis testing that led to the evidence, "*naive*" rater-judges could be used, as in the previous experiment.

If the interviewers did perceive their hypotheses as having been confirmed, the audio and video recordings of the interactions could be analysed by rater-judges to shed more light on which of the two mechanisms caused the effect. If rater-judges who saw the tapes agreed with the interviewers that the targets were behaving in a manner consistent with the hypotheses, then the interviewers really would have made the targets seem to act like the personality profiles in the hypotheses. Alternatively if the interviewers perceived the hypotheses as having been confirmed but the rater-judges saw no difference between the targets in the introvert hypothesis and the extravert hypothesis condition, then the confirmation would have been "*In the eye of the beholder*" only, the first type of hypothesis-confirming bias outlined above.

The rater-judges will see the tapes under two conditions, answers only and both questions and answers. If the rater-judges only see the hypothesis-confirmation under the *answers only* condition, then this would also detract from Snyder & Swann's findings. It would demonstrate that it was only by hiding the situational constraints on the targets' behaviour that Snyder &

Swann were able to force their rater-judges into making erroneous inferences about the target's dispositions.

It should be noted, however, that were no evidence of a self-confirming hypothesis on the part of the interviewers, the judging of the tapes would then be of little theoretical or practical value, and would therefore be omitted due to the time-consuming nature of the empirical work involved.

Medium of Communication

A second factor was also introduced into this experiment to see whether the confirmatory bias was dependent on the impoverished medium of communication used by Snyder & Swann, or whether the effect would be maintained in a full face-to-face interaction. Snyder & Swann (1978) only allowed communication via headphones and microphones. This may have affected the subjects' propensity to bias in person perception, though the literature on media of communication is inconclusive as to whether the audio - only channel of communication is likely to cause more or less error in person perception than the full face-to-face interaction (Short, Williams & Christie, 1976). The possible reasons *why* medium of communication might affect the hypothesis testing process were outlined in more detail in the introduction section of the last chapter.

The crucial measures in this second piece of empirical work were again the number of extravert questions asked in the extravert hypothesis condition compared to the introvert hypothesis condition, and also the interviewers' estimates of the targets' extraversion as measured by the *EPI* and the summated bipolar scales. It was predicted that more extravert questions would be asked in the extravert hypothesis condition than in the introvert hypothesis condition, and thus the opposite would necessarily occur for the introvert questions. It was also predicted that the targets in the extravert hypothesis condition would come to be perceived as more extraverted than the targets in the introverted hypothesis condition by both the interviewers and the rater-judges. No firm predictions were made as to what effect the medium of communication would have on the hypothesis-confirming phenomenon.

If the interviewers' perceptions of the targets reveal that the interviewers <u>have</u> concluded that their hypotheses were confirmed, then the rater-judges' analyses of the interactions would reveal whether there was any basis in the target's behaviour for concluding that the hypothesis had been supported, or whether it was purely an "*in the eye of the beholder*" effect.

Method

Subjects.

Sixty-four female volunteer subjects took part in this experiment. It was conducted at the University of Warwick, but the subjects who took part were a mixture of undergraduates, postgraduates and Open University students. One of each pair was recruited from the library, the other from a coffee bar. This helped to ensure that none of the subjects knew the person they were to interact with more than by sight.

Female subjects were used for reasons of availability and convenience. This was not thought to limit the generalisability of the experiment as the phenonenon has so far seemed to be independent of sex effects.

Procedure.

In the face-to-face condition the two subjects were scheduled to arrive at the laboratory at the same time. A coin was tossed in front of them to allocate them randomly to the roles of interviewer and target. They were then shown into different rooms.

The target completed the extraversion scale from the *EPI* using a computerised version written for this experiment, and was then given the same instructions as

in the previous experiment, telling them that the interviewer would be asking questions of them and they were to answer in as open and revealing a way as possible.

Meanwhile the interviewers were taken through the question-selection task in exactly the same way as in all of Snyder's hypothesis-testing experiments. This differed from the previous experiment in one important way; all of the questions were selected before rather than during the interaction. The 18 questions used were exactly the same as those used in the previous experiment and the interviewers again selected nine.

When this was completed the target was led into the same room as the interviewer, and they sat at opposite sides of the table. They were informed that they would be video-recorded, and the wall-mounted cameras were pointed out to them. The interviewer was then told to start asking her questions when the experimenter had started the recording equipment.

When the nine questions had been asked the experimenter again entered the laboratory, and led the target back to the other room. The interviewer was then asked to use the *EPI* (on the computer) and the ten bipolar scales (identical to those used in the previous experiment) to rate her impression of the personality of the target. While the interviewer did this the target was thanked and debriefed. Finally, the

interviewer was asked to provide a written answer to the question on the last page of the instructions booklet, phrased thus: *Please state whether you felt that the questions that you chose before the interview were suitable for your purpoes, i.e. finding out whether the person you were interviewing was really an extravert (introvert).* They were then thanked and the aims of the experiment were explained to them.

In the audio condition the subjects did not meet before the interaction, and they were given the additional information that the experimenter was interested to see how people got to know each other on the telephone, so they would not see each other. They then communicated from their separate rooms using a microphone and loud-speaker in each room. They were informed that their conversation would be tape-recorded for possible later analysis. Due to an oversight in the preparation of the experimental materials the interviewers in the audio condition were not asked the question about their satisfaction with the questions they had asked. In all other ways the audio and face-to-face conditions were identical.

The design of this experiment was thus a 2 *(introvert hypothesis vs extravert hypothesis)* X 2 *(audio vs face-to-face communication)* fully randomised design, with three dependent variables: the number of extravert questions selected by the interviewer; the interviewers' perception of the target using the *EPI*;

and the interviewers' perceptions of the targets as measured with the summated scales.

The video and audio tapes of the interviews were kept for inspection by rater-judges, but this was conditional upon their being evidence of a confirmatory bias in the interviewers' perceptions of the targets.

Results

The results will be considered in two separate sections for simplicity -- firstly the types of questions selected for asking by the interviewers, then the impressions of the targets formed by the interviewers after they had asked those questions.

The Questions asked.

In order to determine whether the preferential selection of questions seeking confirmatory evidence is replicable with the new list of 18 questions used in this and the previous experiment, a 2 (*Introvert hypothesis* vs *Extravert hypothesis*) by 2 (*audio-only* vs *face-to-face* conditions) analysis of variance was performed on the number of extravert-type questions asked.

The hypothesis factor was significant showing that the average number of extravert questions asked in the extravert hypothesis condition (M=5.75) was greater than the average number of extravert questions asked in the introvert hypothesis condition (M=4.13), $F(1,28)=7.97$, $p<0.005$, one tailed.

Neither the medium of communication effect nor the interaction were significant (both Fs<1). This was to be expected since the treatment of the interviewers in the two medium conditions was identical until after the

question-selection part of their task.

This has shown that the type of questions selecti~~on~~ed is dependent on the hypothesis the interviewer is given to test. Does this bias in the type of questions asked affect the interviewers' judgment?

The impressions formed of the targets by the interviewers.

The same 2 X 2 analysis of variance was performed on the EPI scores given to the targets by the interviewers after the interaction. The mean scores are shown in table 5.1. It can be seen that there is very little difference between the two hypothesis conditions, and the effect was not significant ($F<1$). There was, however, a very weak trend that indicated that targets were perceived to be more introverted in the visual condition than in the audio condition, $F(1,28)=2.55$, $p<0.2$, two tailed.

It is likely that a lot of the variance would have been accounted for by the "actual" extraversion of the target. In the analysis of variance this would have simply been included in the error variance, limiting the power of the analysis. This variance was removed using the targets' self-administered EPI score as the covariate in an analysis of covariance (See Winer, Ch 10 for details).

This achieved more than a fivefold reduction in the magnitude of the error term, making the medium of communication effect highly significant, $F(1,27)=9.90$, $p<0.01$, two tailed. The hypothesis factor was still not significant, the F ratio being a mere 0.02 reflecting not even a hint of difference between the two conditions.

The parallel analyses that were performed for the Summated Scales measure yielded different results. There was no evidence of a medium of communication effect in either the analysis of variance or the analysis of covariance (both Fs<1), but the difference in the means was in the same direction as for the *EPI*, again indicating that targets were perceived to be more extraverted in the audio condition than in the face-to-face condition (Table 5.2).

There was slightly more evidence that the impression formed by the interviewer of the target was effected by the hypothesis, but the difference was nowhere near significant for either the analysis of variance ($F(1,28)=1.78$) or the analysis of covariance ($F(1,27)=2.13$). As can be seen in Table 5.2, the most surprising thing about the difference in these means is that it is in the opposite direction to that which was expected; the targets who were tested for introversion and were asked more introvert-type questions were perceived to be very slightly more extraverted than the

targets who were tested for extraversion!

Neither the EPI nor the Summated scales produced any evidence of a significant interaction (all *F*s<1).

The Interviewer's comments

Fifteen of the 16 interviewers who were asked to report on whether they felt that the questions they chose were satisfactory did so. After reading all of these accounts, several types of comment were noted as being of theoretical interest.

Six (40%) of the interviewers expressed complete satisfaction with the questions. Only one subject complained about the inefficiency of selecting all the questions before meeting the targets. Three interviewers commented in any way that the questions were *leading* or hinted in any other way that the questions were biased. Other causes for dissatisfaction with the questions were that they were too superficial, too probing or too embarrassing to ask. These accounts will be described in more detail in the discussion section.

The Rater-judges.

Since no evidence of an effect for the hypothesis manipulation was found for the interviewers on either the EPI or the Summated Scales, the audio and video

recordings were not analysed to see whether there was any detectable difference in the targets' appearance to "naive" rater-judges.

Table 5.1.
Interviewers' ratings of the targets after the interaction using the *EPI*.

Hypothesis	Communication Medium: Audio only	Communication Medium: Face to face	
Extravert	M = 8.6 SD= 8.1 n = 8	M = 13.1 SD= 6.8 n = 8	M =10.9
Introvert	M = 9.3 SD= 6.8 n = 8	M = 12.6 SD= 6.0 n = 8	M =10.9
	M = 8.7	M = 12.9	

Table 5.2
Interviewers' ratings of targets after the interaction using the Summated Scales.

Hypothesis	Communication Medium: Audio only	Communication Medium: Face to face	
Extravert	M = 19.8 SD= 28.8 n = 8	M = 17.5 SD= 25.8 n = 8	M =18.6
Introvert	M = 31.5 SD= 25.7 n = 8	M = 29.0 SD= 16.4 n = 8	M =30.3
	M = 25.7	M = 23.3	

Discussion

The finding that even when interviewers ask questions that seek out evidence supportive of their hypothesis they still do not automatically accept that hypothesis, goes against all of the assumptions made by Snyder in his reviews of hypothesis-testing in social interaction (Snyder, 1981; Snyder & Gangestad, 1981). This finding and the conclusions that can now be drawn about question selection in interactive settings will be considered seperately.

When will Confirmatory Information-Search occur?

Not only was the "*Confirmatory bias*" effect replicated and found to be highly significant, but also the ratio of extravert questions selected in the extravert hypothesis condition to the number of extravert questions selected in the introvert hypothesis condition was approximately 42:58. This was almost exactly the same ratio as in Snyder & Swann's investigations (if anything slightly stronger, but this may have been caused by the exclusion of the neutral questions). It is reassuring to know that the effect is truly replicable, and particularly with the set of questions used in this experiment.

This means that an important but tentative conclusion from the last experiment can be confirmed; the lack of a difference in the questions chosen to test

the two opposing hypotheses was caused by the questions being selected <u>during</u> rather than <u>before</u> the interaction. This limiting factor on the self-confirming hypothesis effect means that a lot of the interpersonal situations which Snyder postulates would be affected by the phenomenon are in fact unlikely to be affected at all. Certainly, individuals are unlikely to formulate biased questioning strategies in the spontaneous informal conversations that characterise our everyday lives.

Snyder himself seems to have realised in his later writings on hypothesis-testing that the issue of pervasiveness needs to be considered more seriously, rather than simply inferring that hypothesis testing is the framework of all social information gathering. In the paper he co-authored with Gangestad in 1981 Snyder addresses this issue of pervasiveness directly. He concludes that hypothesis-testing is <u>probably</u> occurring whenever an individual is trying to find out about anyone else -- in all situations from the research seminar to the proverbial cocktail party. But they cover themselves by saying even if the predisposition to confirm hypotheses only effects us occasionally in, for instance, employment interviews, then this will still have a major effect on our lives.

The combined findings of this and the previous experiments lead one to conclude that the self-confirming hypothesis could <u>not</u> occur in cocktail

parties or any other informal or spontaneous situation. But what about situations like the job interview, where it is possible that the interviewer has taken time before the interview to prepare a list of questions to ask? The results of this and the other hypothesis-testing experiments lead us to believe that under these circumstances the interviewer would select a majority of questions that would seek confirmation of the hypothesis under scrutiny. Would the asking of these questions then force the confirmation of the hypothesis?

Were the Hypotheses Self-Confirming?

The Interviewers' Perceptions of the Targets.

The answer to this question is again no. Even when the interviewers in this experiment enacted their confirmatory questioning strategy there was no evidence that they were swayed by their questions into believing that the targets' dispositions were more consistent with their hypotheses than was the case. This again is clearly a major blow to the self-confirming hypothesis paradigm. This raises two points of interest: Firstly, why do the interviewers not consider their hypotheses as proven; and secondly, why did Snyder & Swann (1978) arrive at the opposite conclusion?

There are several possible reasons why the interviewers may not have been misled into accepting the target's answers to the biased questions as being evidence supportive to the hypothesis. One possibility

is that the targets simply did not give evidence in favour of the hypotheses in answering the questions. This is unlikely according to evidence from Snyder & Swann's similar experiment (1978, investigation 2) where the naive rater-judges listened to only the targets' answers and found that the targets' responses were consistent with the hypothesis being tested by the interviewer. What is different here is that the interviewers knew which questions were being asked, and thus could compensate for this in assessing the impact of the answer (this point is dealt with in more detail in chapter 3 of this thesis). In particular since most of the interviewers (all bar three to be precise) asked at least three of each sort of question, the interviewers may well have noticed how the targets answered different sorts of questions differently. They may then have avoided simply using an averaging rule to "calculate" the extraversion of the targets, but could have used a more wholistic method to integrate the information from the nine questions. It is, after all, rather simplistic to assume that we use straight-forward averaging rules to deal with the varied and complex information we gather in our social interactions (although some theorists such as Anderson (1971) would argue that we do).

The second issue is why -- when this experiment has shown conclusively that a confirmatory questioning strategy does not cause hypothesis-confirmation -- Snyder and others have argued that hypotheses are self-

confirming.

The actual empirical evidence found by Snyder & Swann (1978) for this claim was examined in detail in chapter 3. To recap briefly, in the original paper Snyder & Swann state that no measure was taken of the interviewers' perceptions of the target after the interview, yet in an article in 1981 Snyder completely contradicts this and says that during debriefing it was noticed that interviewers did think that their hypotheses had been confirmed. Snyder & Swann also claimed that other research had demonstrated that the interviewers would have thought that their hypothesis had been confirmed, but a critical appraisal of this literature showed that this claim was also unjustified. Another mistake has been to assume that because the rater-judges (listening to just the targets' responses) thought that the targets were behaving in a manner consistent with the hypothesis, this was proof that the interviewers would reach the same conclusion. The flawed logic in this argument was also exposed in chapter 3.

So, in fact the empirical findings in this experiment are not in direct conflict with any other empirical findings -- it is the interpretation of those findings that has been the cause of conflicting conclusions. But it is also clear that several features of the methodology used by Snyder & Swann were so artificial that it is difficult to see how any sensible

generalisations to the real world could have been made. What started as the identification of a uniquely social feature of information processing was implimented in experiments that were as far removed from most real-life situations as most purely cognitive psychological experiments.

Medium of Communication.

The only significant effect found for the interviewers' perception of the target was caused by the medium of communication. The interviewers rated the targets as being more extraverted on the EPI when the interaction was conducted in the audio-only condition rather than face-to-face. Why this should have occurred, and why on the EPI but not the Summated Scales is not clear from either this experiment or the mediated communication literature. It is, however, of no direct interest to any of the theoretical issues being considered in this thesis. It was the hypothesis by medium interaction that would have been of more interest, which would have suggested that hypothesis-confirmation is more likely to occur under some situations but not others. There was no evidence of hypotheses being self-confirming in either face-to-face or audio-only interaction.

Rater-Judges

If the interviewers had considered that their hypotheses had been confirmed then it would have been of interest to use rater-judges to explore the process by which this bias had occurred. However, since the role of the rater-judged is of no intrinsic interest in itself, and since the interviewers did not consider their hypotheses to have been confirmed, any data that was obtained from rater-judges viewing the audio and video recordings of the interactions could have added little to the understanding of the confirmatory bias. This part of the proposed experimental procedure was, therefore, not conducted.

The Interviewers' Accounts and the Debriefings

An additional component of the present experiment is that the interviewers themselves were asked whether they were satisfied with the procedures they went through. Were they as sceptical about their task as the empirical findings and critique has suggested they should have been? Reading their written answers to the question they received at the very end of the experiment should shed some light on this.

A casual treatment of the written reports shows that there was very little consensus about any problems the interviewers encountered. Forty percent of the interviewers stated that they were satisfied that the

questions they had chosen permitted them to form a clear impression of the target's extraversion or introversion. Many of the complaints made by the other interviewers concerned issues such as the questions being too superficial and not probing enough or being too probing and embarrassing to ask -- points unrelated to the issues of interest here. Only one subject complained about the artificiality or inefficiency of choosing all of the questions to ask before the interview, and even this was only articulated tangentially by saying that one of her later questions was made redundant by an answer to one of her earlier questions.

One of the specific reasons for asking this question was to see if subjects were dissatisfied with the types of questions they had to choose from. It will be recalled that 21 of the 26 questions on Snyder & Swann's list were of the sort that "would typically be asked of someone <u>already known</u> to be extraverts" or of ".... people <u>already known</u> to be introverts" (1978, p.1204, emphasis in original). Snyder & Swann do not say why the list was made up of this sort of question, rather than questions that set out to <u>differentiate</u> between introversion and extraversion. This and the previous experiment used the same type of questions for the sake of replication -- in fact to an even greater extent by excluding the neutral questions completely. Did any of the interviewers feel dissatisfied with this limited variety of different sorts of questions to ask?

Only a fifth of the interviewers (3 from 15) could be said to have expressed any comments along these lines. One subject said that the questions that she had chosen were ".... *very much loaded for a positive response from an extravert*"; six of the nine questions she had selected were of the extravert type. Another said the questions "... *were not open enough* -- ..." and went on to elaborate on an interesting strategy whereby one might find out whether the target was an introvert by '*"putting them on the spot"*' (her speech marks). This could be done by making them talk a lot and if they became embarrassed or shy one would then know that they were introverted. This could be taken to be an example of a subject considering a falsification strategy (albeit with hindsight -- she had selected seven introvert questions for her initial test of an introvert hypothesis). The third subject, who chose all nine extravert questions to test the extravert hypothesis could maybe be considered to have criticised the types of questions by calling them "restricted", and she stated that whereas she had originally chosen "*questions which were associated with being lively*" she now thought that it would have been more valuable to also enquire about times that the target was introverted.

What is clear from this is that only a few subjects criticised the questions for being of this "... <u>already known</u> ..." sort when asked specifically about the questions, and even then it was apparent that of these

individuals some had only come to this realisation after asking their questions. This is not to say, of course, that given a free choice, they would generate this sort of question themselves. That is to be the topic of the next experiment.

Conclusions

1/ The interviewers selected questions that sought confirmation of their hypothesis. This proves that the absence of the "confirmatory bias" in the previous experiment was not an artefact of the particular list of questions used, but was caused by the interaction of the interviewer and target. This finding severely limits the external validity of any confirmatory bias in the testing of hypotheses about other people in social interaction.

2/ Even when the interviewers did select and ask questions that sought confirmatory evidence this did not necessarily lead them to conclude that their hypotheses had been supported. This clearly undermines the effect that Snyder claims to have demonstrated, but it is not clear whether it is, in fact, contrary to any actual empirical findings. Snyder (1981, Snyder & Swann, 1978) gives contradictory accounts of what he actually did and did not measure.

3/ Subjects were generally satisfied with the types of questions they were given to select from and

with the selection procedure. The few subjects who did comment on the inadequacies of the questions or procedure seem to only have noticed the inadequacies after selecting their questions and asking them.

Chapter 6
Do People Really ask Biased Questions?

Introduction

One feature of the experimental paradigm that has not been explicitly challenged in the two previous experiments is the sort of questions that the interviewers were given to choose from.

A conceptual analysis of the type of questions suggests that the choice of questions that Snyder & Swann first used in 1978 and others copied (Snyder & Campbell, 1980; Snyder & White, 1981; Cooper, 1982 and Semin & Strack, 1980) was not very well thought out.

To recap, Snyder & Swann's subjects were required to select 12 questions from a list of 26 compiled by the experimenters. Of this list "*.... 11 questions were ones that the majority of rater-judges thought would typically be asked of people* <u>*already known*</u> *to be extraverts, ... 10 questions (that) would character-istically be asked of individuals* <u>*already known*</u> *to be introverts ..... (and) ..... 5 questions for which there was no consensus that they were extraverted questions or introverted questions and those ..... irrelevant to introversion or extraversion*" (1978, p.1204, emphasis in original).

However, the most appropriate questions to ask for the interviewers, given that their task was to differentiate between introverts and extraverts, would not have been any of these sorts of questions. At no point does Snyder ever question or justify this choice of questions. In all of his examples and anecdotes he just assumes that these are the sort of questions suitable to the testing of hypotheses.

On the other hand, there seems to be little evidence that subjects were dissatisfied with the choice they were given. In the last experiment only three of the 16 subjects expressed dissatisfaction with the questions, and that was when they did not even have the third category of neutral questions to choose from.

Another sign that the subjects did not feel unhappy about choosing from the list can be obtained from the fact that when they did have a neutral category to choose from, the subjects chose those questions slightly less frequently than one would expect by chance. For example, on average over investigations 1, 2, and 4 of Snyder & Swann's (1978) experiment subjects chose to ask an average of 1.94 neutral question, that is 39% of the total number available. This compares to 10.04 of the combined introvert and extravert questions that were chosen, or 48% of the total of 21. This argument could be countered, though, by the fact that some of the five neutral questions were possibly irrelevant to introversion or extraversion (on topics such as favorite

charities).

This experiment is aimed at finding out what sort of questions individuals would generate themselves in order to test a hypothesis, if they were not constrained to choosing from a list.

The procedure to be used in this experiment is thus very simple and very similar to the procedure used in the last experiments. The main difference is that the subjects, instead of being given a list of questions to choose from, were simply told to write down the questions they wanted to ask.

Two main hypotheses can be proposed. Firstly, it is not clear why subjects should assume, in the wording of their questions, that the targets should already possess one of the traits they are testing for. For example, instead of asking a question like "*What do you dislike about loud parties?*" it would seem more sensible to ask the simpler question "*Do you dislike loud parties?*". Thus when the questions are categorised by judges, fewer of the "*subject generated*" questions should fall into the category of questions that make unjustified assumptions about the target than the proportion of such questions on Snyder & Swann's list.

Secondly, though, there may still be some evidence of a "confirmatory bias", albeit in a much weaker form.

The subject matter that interviewers use in the phrasing of questions could either be drawn from the activities and feelings of a typical extravert, or they could be drawn from the activities and feelings associated with introversion. If subjects are simply using the personality profiles to provide topics to ask questions about, then it would be expected that subjects would ask more questions about introverted topics when testing for introversion, and ask more questions about topics associated with extraversion when testing for extraversion.

This experiment will also be used to investigate a further aspect of the confirmatory bias. It was noted before that one possible expla nation of why individuals tend to choose questions that seek out confirmation of hypotheses originates in the way that hypotheses are phrased. The personality profiles given to the subjects gave plenty of examples of the type of evidence that would confirm the hypotheses, but no examples of the types of evidence that would weaken or falsify the hypotheses. Thus one way to weaken or remove the confirmatory bias would be to make the interviewers aware of evidence that would run counter to the hypothesis while they are choosing their questions. One attempt to do this was by rephrasing the personality profiles so they contained both evidence that would support and evidence that would contradict the hypothesis (Snyder & Campbell, 1980). This attempt, in fact, failed to detract at all from the power of the

confirmatory bias -- just as many confirmatory questions were asked as when the hypothesis was framed in exclusively confirmatory terms. An even more extreme manipulation was performed in an unpublished experiment in which the profiles were framed exclusively in terms of the characteristics that would be absent in someone of that personality type. Again the subjects continued in their attempt to use the questions to solicit confirmatory evidence (Snyder & White, 1978).

An independent line of enquiry in a separate realm of psychology also concluded that the way in which logical tasks are phrased can be crucial in determining the way individuals select information to test hypotheses. Evans (1972) was exploring the parameters of subject's defective reasoning in Wason & Johnson-Laird's (1972) propositional reasoning tasks. One way he found of manipulating the solutions given by subjects was to alter the phrasing of the propositions, from the "<u>if p then q</u>" form to "<u>if not p then not q</u>", "<u>if p then not q</u>" or "<u>if not p then q</u>". There was a consistent bias in the subjects' proposed solutions to the task to match rather than to alter the values named in a rule.

The reason given by Snyder & Campbell for their failure to remove the effect was that perhaps the notion of introversion-extraversion is so well known and familiar to the lay-person that the simple change of phrasing of the personality profile was not powerful enough a manipulation to change the subject's mental

representation of the construct. Another way of making the subjects aware of the alternative to a hypothesis would be to give them not just one description of the evidence required to confirm the hypothesis, but also a separate description of the types of evidence that would be needed to falsify the hypothesis. This may sensitise them to the importance and relevance of falsifying evidence more than by simply providing a mixture of examples of falsifying and confirming evidence. Therefore a second, orthogonal manipulation to the hypothesis factor was the number of personality profiles that subjects saw. In the "*confirmation-only*" condition subjects would see just the description of the personality type that they were testing for. In the *"confirmation and disconfimation*" condition the subjects would receive both the introvert and the extravert personality profiles to read before they prepared their questions to ask. The hypothesis was phrased in exactly the same way in both conditions, asking the subjects to test to see whether the person they were to interview was of the type in the <u>extravert</u> (or, in the introvert hypothesis condition, <u>introvert</u>) description.

If this manipulation does have the predicted effect, then there will be an interaction between the two independent variables for the introverted and extraverted questions such that the confirmatory bias is weaker or non-existant in the "both profiles" condition than in the "confirming profile only" condition. Also, it is likely that there will be more neutral questions

in the two-profile conditions than in the one-profile conditions.

The three hypotheses can be summarised thus:-

Hypothesis 1

The questions generated by subjects will contain very few of the questions used predominantly by Snyder & Swann (1978) -- that is, questions one would ask of people <u>already known</u> to be either extraverted or introverted.

Hypothesis 2.

There will be some difference between the questions generated to test the two hypotheses. Subjects will tend to ask more questions about introverted behaviours or feelings when testing the introvert hypothesis, and more questions about features of extraversion when testing the extravert hypothesis.

Hypothesis 3.

The weaker form of the confirmatory bias outlined in Hypothesis 2 will be weakened further by giving the subjects a personality profile of the opposite personality to the one they are testing for, as well as the personality profile of the typical individual who would confirm the hypothesis under test. There will be less of an imbalance between introvert and extravert questions, and more neutral questions when both personality profiles are presented.

Method

Subjects

Twenty male and 20 female undergraduates from the University of Warwick and the Open University volunteered as subjects. In addition two male postgraduate psychology students, naive to this program of research, were used as rater-judges.

Procedure

The procedure for this investigation closely followed the procedural paradigm developed by Snyder & Swann, (1978) and used in the other hypothesis-testing experiments in social psychology. Participants were told to prepare to interview another undergraduate to determine the extent to which the target's behaviour and experiences match those of a typical extravert (extravert hypothesis condition) or those of a typical introvert (introvert hypothesis condition) They were told that they should prepare nine questions to ask of that person, and that they should write those questions on the sheet provided.[1]

The second factor was introduced by giving some subjects just one personality profile to read; the subjects in the other condition received two personality profiles, which included the opposing personality profile as well as the profile they were testing for.

When the subjects had written down their nine questions they were told that, in fact, no interview would take place. The reason for the deception was explained to them, and they were debriefed and thanked.

This part of the experiment formed a 2 (*introvert hypothesis* vs *extravert hypothesis*) X 2 (*one profile* vs *both profiles*) fully randomised design. The dependent variables were to be the frequency with which the "interviewers" generated various types of questions. Two rater-judges were used to determine this.

Categorising the questions

The 26 questions from Snyder & Swann's list were combined with the nine questions from each of the 40 subjects, making a total of 386. They were mixed in a randomised order and typed in a format to make Snyder & Swann's questions indistinguishable from the questions generated by subjects themselves.

The two rater-judges were then individually and carefully briefed on the categories they were to allocate each question to. It was stressed that while the categories were not necessarily mutually exclusive, it was important to assign each question to the one category that it fitted most closely.

The five categories were described thus:-

Extravert Questions which ask about a feature derived from the extravert personality profile (Refered to from now on as "E" questions for brevity).

Introvert Questions which ask about a feature derived from the introvert personality profile ("I").

Biased extravert questions which assume that the target is an extravert and do not allow an answer indicative of introversion ("BE").

Biased introvert questions which assume that the target is an introvert and do not allow an answer indicative of extraversion ("BI").

and

Neutral or Non-directional questions which either present a choice between an extravert and an introvert alternative, or are open ended, or are irrelevant to the introversion-extraversion dimension ("N").[2]

The rater-judges were also given the following examples of each type of question to clarify the categories:-

E: *Do you like parties?*

I: *Do you get nervous in the company of strangers?*

BE: *Why do you like socialising so much?*

BI: *In what situations do you find it impossible to talk to strangers?*

N: *What are your hobbies?*

or

Do you prefer reading books or talking to people?

or

What sort of music do you like?

The two judges worked through the list of questions seperately in their own time. A full set of instructions, including the two personality profiles they were each provided with, is included in Appendix 6.1

1 A pilot test had revealed that subjects had considerable difficulty in thinking up 12 questions to ask, so the number was reduced to nine.

2 The categories used here were a close copy of the categories used in an unpublished paper by Trope & Alon (1980). Several other methods of categorisation were attempted, for instance measuring the degree of constraint put on the answer by the question on a 7-point scale. Although the overall pattern of results was the same regardless of the categorisation method tried, the inter-rater reliability was found to be much lower with other methods.

Results

The rater-judges agreed on the categories of 70.2% of the questions[1]. Cohen's Kappa was computed as 0.53 ($p<0.001$), showing that the agreement between rater-judges was far higher than one would expect by chance alone.

The information from the two rater-judges was combined to compute an average number of questions in each category for each subject. For instance, if one rater-judge had put four of a particular subject's questions into the neutral category, and the other judge had classified five as neutral, then that subject would score 4.5 for the neutral category.

The proportion of subject-generated questions that fell into the two biased categories compared to the proportion of Snyder & Swann's questions that were categorised as biased is shown in Figure 6.1. It can be seen that, as predicted, the proportion of biased questions generated by subjects (2.8%) is far less than the proportion on Snyder & Swann's list (50%)[2]. A chi-squared test showed this difference to be highly significant ($X^2=185$, df=1, $p<0.001$).

In order to test the second and third hypotheses the numbers of introvert, extravert and neutral questions generated by each subject were analysed seperately using a 2 (introvert hypothesis vs extravert

hypothesis) X 2 (one personality profile vs both personality profiles) analysis of variance. There were too few biased questions generated to satisfy the assumptions of a parametric test, but they were asked with approximately the same frequency in all conditions.

For the introvert questions there was a significant difference in the types of questions asked in the two hypothesis conditions; more introvert type questions were generated to test the introvert hypothesis (M=1.95) than to test the extravert hypothesis (M=0.82), $F(1,36)=6.05$, $p<0.01$, one tailed). (Table 6.1).

The difference in the means for the extravert questions was also in the predicted direction, with more extravert questions being generated to test the extravert hypothesis (M=2.10) than to test the introvert hypothesis (M=1.97), but this difference was not significant ($F<1$). (Table 6.2)

One difference was found due to the manipulation of the "*confirmatory only*" compared to the "*confirmatory and disconfirmatory*" personality profiles. Although there was no support for the third hypothesis from the introvert or extravert questions (in both cases *Hypoth-esis* by *Profiles* interaction Fs<1), there was clear support for it from the neutral and non-directional questions (Table 6.3). More neutral and non-directional questions were generated when subjects were given both personality profiles with the hypotheses to test

(M=5.875) than when only the confirmatory personality profile was presented to them (M=4.77), $F(1,36)=3.12$, $p<0.05$, one tailed.

1 This is much lower than the figure of 92% given by Trope & Alon There are two possible reasons for this. Firstly, the conditions under which the subjects in Trope & Alon's experiment may have been different -- it seems as if they were encouraged to ask questions that required a simple yes/no response, which were probably easier to categorise. Secondly, the inter-rater reliability is a function of the similarity between raters. Whereas the raters in this study were chosen for their different areas of study, the raters in Trope & Alon's study may have been more homogeneous.

2 This figure is seemingly at odds with Snyder & Swann's judges who put 21 of the 26 questions, or 81%, in their equivalents of the biased categories. This difference is, however, entirely attributable to the fact that there were only three possible categories available to those judges, as opposed to five available to the raters in this experiment.

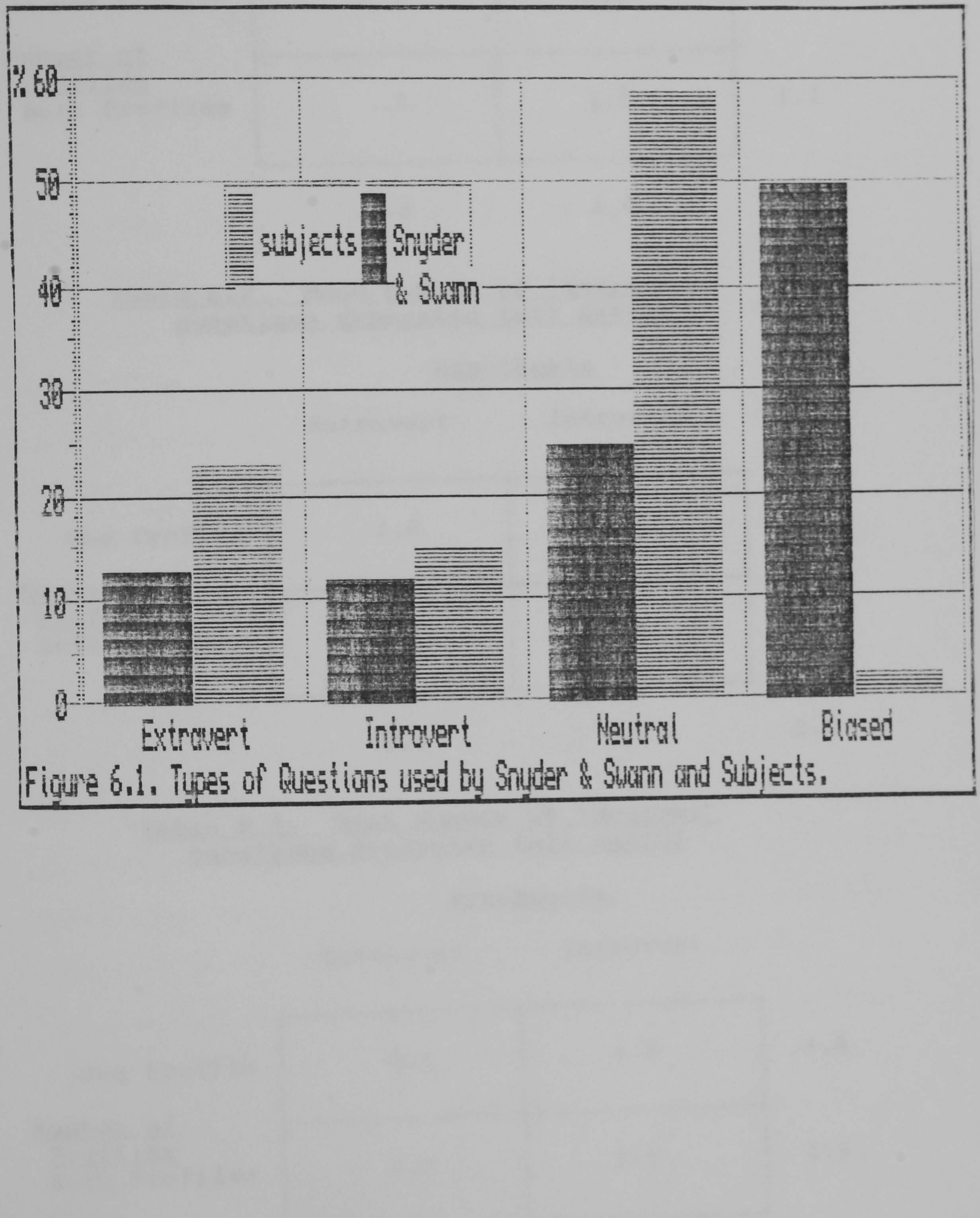

Figure 6.1. Types of Questions used by Snyder & Swann and Subjects.

Table 6.1. Mean Number of "*Introvert*" questions generated (all *n*s=10)

Number of profiles	Hypothesis: Extravert	Hypothesis: Introvert	
One Profile	1.1	2.2	1.7
Both Profiles	.6	1.7	1.1
	.8	2.0	1.4

Table 6.2. Mean Number of "*Extravert*" questions generated (all *n*s=10)

Number of Profiles	Hypothesis: Extravert	Hypothesis: Introvert	
One Profile	2.5	2.1	2.3
Both Profiles	1.9	1.8	1.8
	2.1	2.0	2.0

Table 6.3. Mean Number of "*Neutral*" questions generated (all *n*s=10)

Number of Profiles	Hypothesis: Extravert	Hypothesis: Introvert	
One Profile	5.1	4.5	4.8
Both Profiles	6.5	5.3	5.9
	5.8	4.9	5.4

Discussion

The three hypotheses will be considered seperately before their combinet impact on hypothesis-testing in social interaction is considered.

Hypothesis 1 received overwhelming support. While the judges rated 50% of the questions on Snyder & Swann's list as being in one of the biased categories, less than 3% of the questions generated by subjects were judged to be of this sort. This meant that most subjects would not have generated any of these sort of questions at all. The logic of using the biased questions on the list is now very suspect. It was clearly a mistake, and it is dubious whether any experiments using this type of questions can have any direct bearing on real-life social interactions.

Hypothesis 2 received firm support from the introvert questions, but only weak non-significant support from the extravert questions. An inspection of the means shows that over twice as many introvert-type questions were asked to test the introvert hypothesis compared to the extravert hypothesis. The other interesting feature is the very low overall frequency with which introvert questions were asked in either condition. Even when testing the introvert hypothesis subjects generated an average of less than two introvert questions, less than the number of extraverted questions they generated.

The reason for this lack of difference in the mean number of extravert questions asked in the two conditions may be a function of the way in which the introvert-extravert dimension is represented linguistically. Extraversion is thought to be the "positive" end of this dimension. To ask how outgoing or extraverted somebody is is a fairly neutral question, but to ask a question like "How shy or introverted is the new member of staff?" is a "*loaded*" question that strongly suggests that there is reason to believe that she is, to some degree at least, introverted. Thus the mention of "introversion" probably calls "extraversion" to mind automatically, whereas the mention of extraversion does not necessarily make the notion of introversion as available. (Clark & Clark, 1977).

Another possible expla nation is that the introvert questions were less "socially desirable" than the extravert questions. Cooper (1982), using a list of questions similar to those used by Snyder & Swann (1978), asked subjects to rate how comfortable they would feel asking each question. Subjects consistently rated the introverted questions as being less easy to ask, which Cooper concluded was the reason that fewer introverted questions than extraverted questions are selected to be asked in all of these hypothesis testing experiments. This can be seen from a quick look at the list of questions; the extravert questions are about

events with connotations of positive affect like making friends and going to parties, but the introvert questions are about failures in social functioning, marked by feelings of lonliness, shyness and awkwardness in dealings with others. This would account for not only the fact that fewer introvert questions were asked overall, but could also account for the fact that so very few were asked in the extravert hypothesis condition. If subjects are asked to test for introversion they may feel that they have little choice but to ask at least some questions about introversion, but if they are told to test for extraversion they not may feel so compelled to ask those "less friendly" questions. There is, after all, strong evidence that the impression that the interviewers think that the targets will gain of them is a very powerful influence in the selection of questions (Snyder, Campbell & Preston, 1982).

Hypothesis 3 also received some support. More neutral questions were asked in the *two-profile* condition than the *one-profile condition*. These neutral questions are probably the fairest and best sort of questions to ask to test between two competing hypotheses because they either allow a direct comparison between a feature of extraversion and a feature of introversion, or they allow the targets to choose their own "arena" for their answer. The debriefing of the subjects revealed one further tactic in asking questions that fell into this category; to

ask questions that would make the target talk, and in doing so their extraversion could be gauged from their mannerisms and style rather than the actual content of what they said.

The implications for hypothesis-testing in social interaction.

The conclusions drawn by Snyder that any hypothesis testing that takes place in social interaction would be doomed to automatic confirmation (Snyder & Swann, 1978; Snyder, 1981) now look even more unfounded. As well as the evidence amassed so far about the artificiality of choosing questions to ask before rather than during an interaction, and as well as the conceptual criticisms of some of Snyder's methodology, and as well as the demonstration that interviewers do not consider their hypotheses as having been confirmed, there is now clear evidence that the questions used in all of the previous studies were very different from those actually generated by individuals to test hypotheses anyway. It is difficult to imagine how the sorts of questions generated here could ever lead to self-confirming hypotheses; they would seem to give the targets complete freedom to answer in a manner indicative of their true personalities, and without constraining them into accepting false premises.

One point of interest to emerge is the way in which the vast majority of subjects fully accepted the list

of questions they were given to chooose from in previous experiments, despite the fact that they would hardly ever generate that sort of question to ask themselves. It would be of interest to find out to what extent individuals are aware of the constraints that some types of questions can put on the respondant.

This experiment seems to have highlighted what Cohen (1981) would call a "*cognitive illusion*". Subjects have been made to behave in a seemingly irrational way in the laboratory, but only because of the grossly artificial situation which the experiment has put them in. With a few moments' reflection, or with a little help, the subjects would realise that what they were doing was not rational and correct it.

But, would there be any time when subjects would ask "biased" questions? One situation in which subjects might be induced to ask questions that already assume that a person is an introvert or an extravert is if they had a plausible reason to make that assumption. What would happen, then, if subjects did have a good <u>a priori</u> reason to believe that their hypothesis was likely to be true? This is the main empirical question set for the next experiment to investigate.

Another unanswered and important question is how individuals phrase the hypotheses they set themselves. If subjects ask more neutral questions when given a hypothesis in terms of two competing alternatives

rather than just one possibility, then it becomes important to know the exact form of the hypotheses individuals set themselves to test in their social lives. If they are of the form "*I wonder whether the new member of staff is an extravert or an introvert?*", rather than "*I wonder whether the new member of staff is an extravert?*" then there is even less reason to think that questioning strategy is going to be a function of the hypothesis under test. This point will be discussed again in more detail in chapter 9.

Conclusions

1/ When choosing their own questions to ask, individuals hardly ever choose *biased* questions. Instead they ask questions that both introverts and extraverts could answer equally well.

2/ It is very difficult to see how the asking of these "*fair*" questions could lead to self-confirming hypotheses. This makes most of the generalisations from the experiments that have used Snyder & Swanns (1978) list of questions totally unwarranted.

3/ There is a slight tendency for subjects to still draw upon features of the hypothesis that they are testing rather than its alternative in formulating questions to ask. While this could be called a very weak form of a *confirmatory bias* it probably could not lead to interviewers confirming their hypotheses.

4/ In any case, the majority of questions generated are of the *neutral* sort, which either contrast features of the hypothesis and features of the alternative or are completely open ended allowing the target to answer in an unconstrained way.

5/ When presented with personality profiles of both the typical person who would conform to the hypothesis and the person opposite to the hypothesis, subjects were even more likely to formulate *neutral* questions. This further weakens any predisposition to think in terms of the hypothesis rather than the alternative. It is not known whether people naturally think in terms of single hypotheses rather than competing pairs. The research to date has, without justification, always assumed that people think solely in terms of the hypothesis and not the alternative. If this is not the case it is yet another flaw in the self-confirming hypothesis paradigm.

6/ The findings of this experiment, taken with the two previous experiments, make self-confirming hypotheses seem even less likely to occur in social interaction.

The next experiment is an attempt to find out if interviewers might use biased questions if they had good reason to believe that their hypotheses are most likely to be true.

Chapter 7

Are *Biased* questions asked to test high-certainty hypotheses?

Introduction

The last experiment showed that under normal conditions of hypothesis testing in social interaction, individuals do not generate questions that could bias targets into answering in a manner representative of introversion or extraversion. But are there any conditions in which interviewers might formulate "biased" questions of the type used by Snyder & Swann (1978)?

One possible situation which may give rise to this would be if the hypothesis testers had reason to believe that the hypotheses they were testing was probably true. Under these circumstances individuals may perceive hypothesis testing as being a task which involves collecting a few more bits of evidence to become certain that the hypothesis is true, rather than an open-ended quest for information.

Snyder & Swann manipulated the certainty of the hypothesis in two of their investigations, to see whether it affected the questioning strategies of the interviewers. In their first investigation they manipulated the alleged origin of the hypothesis.

Subjects in the "*Low Certainty*" conditions were not given any information as to the origin of their hypotheses, in the same manner as in the three experiments described so far in this thesis. Conversely, subjects in the "*High Certainty*" condition were told that their hypotheses were derived from the results of a personality test taken by the target the previous week, and it was their task to see whether the results of the test were accurate by looking at the behaviour of the target herself. Pretests had shown that undergraduates had considerable faith in personality tests, so Snyder & Swann reported that they were confident that this was an effective and strong manipulation of the certainty of the hypothesis.

Snyder & Swann reported that the "*confirmatory bias*" was unaffected by this manipulation. The interviewers were just as keen to use their questions to search for confirmation of their hypotheses regardless of whether they thought that the hypothesis was highly likely to be true or whether the hypothesis was merely a hypothetical premise (1978, investigation 1).

The third investigation went one stage further in attempting to encourage the subjects to reject their confirmatory strategies. In this experiment all of the subjects were instructed to test for extraversion, but some were given a compelling reason to believe that their hypotheses were probably true, and others were given a compelling reason to think that their hypotheses

would probably turn out to be false. This was done by telling the interviewers that the individual they were to interview was in a sorority of which all the members had been previously tested for extraversion. The interviewers in the "*Many extraverts*" condition were informed that the results had shown that 23 of the 30 members were extraverts, so it was their task to find out whether the particular member they were to interview was one of those many extraverts. In the "*Few extraverts*" condition the interviewers were told that only seven of the 30 members of that particular sorority were extraverts, and their task was to find out whether the particular member they were to interview was one of those few extraverts. As before, a manipulation check ensured that subjects were fully aware of the implications of the instructions.

Snyder & Swann (1978) concluded, from their analysis of the questions selected to test the hypotheses in these two conditions, that the strength of the "*confirmatory bias*" was unaffected by manipulating the certainty of the hypothesis. Even when subjects thought that the person they were to interview was <u>unlikely</u> to conform to the hypothesis, they still selected questions that searched for confirmation of that hypothesis. This analysis compared the number of extravert, introvert and neutral questions asked under the "*few extravert*" and "*many extravert*" conditions with "comparison conditions" constructed by adding together various sets of data from investigations 1 and 2.

A re-analysis of Snyder & Swann's results

Given the general inadvisability of comparing results from one experiment directly with results from other experiments, it is difficult to understand why they went to such lengths to analyse two groups of data. An analysis that would have been both simpler and more powerful could have been performed by three *t*-tests between the number of extravert, introvert and neutral questions asked in the "*many extraverts*" as opposed to the "*few extraverts*" conditions. This re-analysis reveals that significantly more extravert questions were selected for asking in the "*many extraverts*" condition (M=6.93) than in the "few extraverts" condition (M=6.07), $t(28)=1.80$, $p<0.05$, one tailed. There was a corresponding trend, just short of statistical significance, for more introvert questions to be chosen to test the extravert hypothesis under the "*few extraverts*" condition (M=3.87) compared to the "*many extraverts*" condition (M=3.00), $t(28)=1.53$, $p<0.1$, one tailed. Exactly the same number of neutral questions were selected in both conditions (Ms=2.07). An inspection of the data from the condition of interest, where interviewers prepared to test a hypothesis that they thought was probably false, reveals that the behaviour of the interviewers still resembles the "*extravert hypothesis*" questioning strategy much more closely than the strategy adopted to test the introvert hypothesis. The correct interpretation of the

results seems to be that while interviewers will still seek confirmation for a hypothesis they think is likely to be false, there is also evidence that this effect is weaker when there is a negative expectation.

Question types and certainty of hypotheses

In light of the results of the previous experiment, though, it seems as if Snyder & Swann's may have been looking in the wrong direction for the effects of certainty of the hypothesis. Instead of asking more or less of the same types of questions depending upon the certainty of the hypothesis, individuals may ask different types of questions depending on whether they consider the hypothesis more or less likely to be accurate. The results from the previous experiment demonstrated that when the interviewers had no compelling reason to suggest that the hypothesis was accurate, they ask only an exceedingly small proportion of "*biased*" questions. If one thing is likely to make them generate questions that do make assumptions about the target, it may be if they were given information to lead them to believe that the hypothesis is probably true. They may then see their task of finding out exactly how much of an extravert the target is, instead of whether he or she is an extravert or an introvert. When distinguishing between levels of extraversion, the interviewer may well inadvertantly assume at least some degree of extraversion. This may be the foundation of self-confirming hypothesis in social interaction.

Examples of some of the biased questions used by Snyder & Swann (1978) will help to illustrate this point. In testing for introversion, instead of asking whether the target disliked loud parties, the interviewer may ask "What things do you dislike about loud parties?" if she has strong reason to believe that the target is an introvert. Similarly, if the interviewer thought that the target was very likely to be an extravert, she might be tempted to ask "What do you like about living in situations in which there are lots of people around?" instead of simply enquiring about preferences in living arrangements. Both of these questions (from Snyder & Swann's list) are of the sort that would make most people relate their introverted and extraverted behaviours respectively, regardless of their actual personality.

This experiment is designed to look more closely at this possibility. If the theoretical analysis above is correct then one would expect more of the "*biased introvert*" or "*biased extravert*" questions to be selected to test hypotheses that have a high subjective probability of being true than when testing hypotheses with no information to suggest that they are true.

The present experiment also presents the opportunity to test several further hypotheses. Firstly, this experiment will act as a constructive replication of the findings of the previous experiment

that the proportion of biased questions selected to test hypotheses with no evidence to suggest that they are true will be very small. It can also be used as an opportunity to replicate the weak form of confirmatory bias found in the last experiment, whereby more introvert type questions are selected to test an introvert hypothesis than an extraverted hypothesis, and more extravert type questions are be selected to test extravert hypotheses than introvert hypotheses. This experiment will return to the methodology whereby the interviewers select questions from a list in preparing to test a hypothesis, rather than the form of the task in which the interviewers generate their own questions. It will then be possible to see whether any of the differences between the questions selected from a list in the experiments reported in Chapters 4 and 5 and those generated by the subjects in the last chapter were simply artefacts of the method of selection.

Social Desirability

A final test of interest was to see why, in all of the hypothesis testing experiments tested here and elsewhere, there are always more extravert questions asked than introvert questions. It is possible that the biased introvert questions from Snyder & Swann's list are more embarrassing to ask of strangers, particularly ones likely not to be introverts (this point was discussed in more detail in the previous chapter). Rater-judges will rank all of the questions on a scale

according to how comfortable they would be to ask of a stranger. Two hypotheses can be proposed from these scores. Firstly it is predicted that the judges will report that they would be less comfortable asking the *biased introvert* questions than any other category. Secondly it is predicted that there will be a positive correlation between a question's "*comfortableness*" rating and the number of times it is selected by the "*interviewers*".

Method

Subjects

Twenty male and 20 female subjects volunteered to participate in this experiment. They were all attending an Open University summer school at the University of Warwick, studying either introductory Psychology or technology.

A further 8 undergraduates and postgraduates from the University of Warwick were recruited to act as rater-judges.

Materials. The Question list.

This experiment required a list of questions for subjects to choose from, which had to be constructed. It contained 48 questions, eight from each of the following categories:-

Biased Extravert

Biased Introvert

Extravert

Introvert

Neutral

and

Irrelevent.

The questions for each of these categories were selected from the questions categorised by the judges in the previous experiment. The same rater-judges re-categorised the neutral questions to form a seperate category for any questions that they considered irrelevent to introversion-extraversion. All of the questions that the judges had not placed in the same categories were excluded, and eight of the remaining questions of each type were selected randomly. A check was made to eliminate any questions that were very similar in form and content to other questions. One of such pairs of questions was replaced with another question from the same category. The 48 questions were then put in a random order and presented on three typed A4 sheets, entitled "Questions used in Interviews". A copy of this list is included in Appendix 7.1.

Procedure.

When the subjects arrived at the laboratory they were told that this was an experiment on impression formation. They were given written instructions that informed them that they were going to interview another person to determine whether they were an extravert (introvert). As in previous experiments they were told to select questions to ask from a list provided by the experimenter. They were also given the personality profile with the description of a typical extravert (introvert), and the list of 48 questions. They were told to read the personality profile and then study all

of the 48 questions carefully before selecting the questions to ask. A pretest had revealed that subjects found that selecting questions from the longer list was fairly difficult and time consuming, so the number of questions to be selected was reduced to eight.

The certainty of the hypothesis was manipulated by the information given relating to the origin of the hypothesis. In the "*hypothesis plus expectancy*" condition, the interviewers were told that "*a personality test taken by the person concluded that their characteristics were very similar to the description in the profile; we want you to see whether they actually come across like that when they meet someone.*".

In the "*hypothesis only*" condition interviewers were simply told "*Your task is to find out how well the introvert profile describes the person you interview.*".

In both conditions the rest of the instructions were similar to those given by Snyder & Swann (1978, investigation 1) to their subjects.

The subjects were left alone to select their eight questions by ringing the numbers of those questions on the list. They were told to inform the experimenter when they had done this, so they could be introduced to the person they were to interview. The selection of the questions typically took about 10 to 15 minutes. After

this time the subjects were informed that, in fact, the interview was not going to take place but that it was the selection of the questions themselves which was the focal point of the interview. The selection of the questions was discussed in detail with the subjects before the aims of the experiment were explained to them. The experimenter apologised for the deception used but explained why it was necessary, and the subjects were thanked for participating.

The design of the experiment was fully randomised with two factors and two levels to each factor, high vs low certainty of the hypothesis by introvert vs extravert hyothesis. There were ten subjects per cell, five males and five females.

In addition to the main experimental procedure, one other measurement was included to aid in the interpretation of the results. The eight rater-judges were asked to read through the list of 48 questions, and rate each of the questions on "*how comfortable you would feel asking that question to a person you had not met before?*" They were to rate each question on a seven point scale from one "*I would feel very uncomfortable indeed*" to seven "*I would not feel at all uncomfortable*". They were carefully briefed by the experimenter before completing the task individually in their own time -- usually about 20 minutes to half an hour.

Results

The analyses of the data were different for each of the hypotheses being tested, the only common element being that the number of questions asked from some or all of each of the six categories formed the dependent variables.

For the reasons already discussed in chapter 6 comparisons of the frequencies of each type of question asked with each other would be very difficult to interpret. This is because no attempt has been made to control for confounding variables including differences between the questions other than those directly relevant to the criteria of categorisation. For instance, some of the questions may happen to be more diagnostic or attractive for whatever reason -- no attempt has been made to control for this. However, the approximate frequency with which the various categories are selected will be of interest, given the exceedingly large disparity between frequencies found here compared to the last experiment.

Are *biased* questions more likely to be asked in the *hypothesis plus expectation* conditions?

The main hypothesis was that more biased questions would be asked to test the hypothesis when it was accompanied by a positive expectation. Evidence for this would be found by two planned t-tests, for the

biased extravert questions between the two levels of the certainty factor within the extravert hypothesis, and for the biased introvert questions within the introvert hypothesis. Tables 7.1 and 7.2 show the mean number of each type of question asked in each of the four cells. The t-test for the biased extravert questions provided support for the hypothesis with a difference in the means that was just significant. Subjects testing the extravert hypothesis selected more biased extravert questions to ask if they had some evidence that the hypothesis was probably true (M=2.4) compared to the subjects in the "hypothesis only" condition (M=1.4), $t(18)=1.91$, $p<0.05$, one tailed. The difference in the introvert questions was, however, in the opposite direction to that predicted. More biased introvert questions were asked when the subjects testing the introvert hypothesis had no reason to believe the hypothesis (M=0.9) than when they had some firm evidence supportive of the hypothesis (M=0.8)! Since one-tailed tests were planned and the difference was in the opposite direction to that predicted, no inferential test was employed but it can be seen that the difference between these means is, in fact, negligible.

Were *open-ended* questions prefered to *biased* ones?

The second hypothesis was that the number of biased questions selected would be very small. This was not the case; The combined total of biased questions was an

average of 2.6 questions out of eight, or 32%. This was almost exactly the same number that would have been chosen if subjects were selecting by chance alone (33.3%), so it does not seem as if subjects were attempting to avoid bias in their questions. A goodness of fit chi-square test showed the high degree of closeness to this chance level, $X^2=0.3$, df=1, $p>0.8$. A further surprise is that even the *irrelevant* questions were selected fairly frequently, 12% of all questions asked being from this category compared to a "chance guessing" figure of 16.7% (see Figure 7.2).

One valid comparison that can be made is with the subjects in the last experiment who selected only 2.8% biased questions. A Chi-square test between these two experiments shows that the proportion of biased questions generated in the present experiment (32%) was clearly much higher than the proportion selected in the last experiment (28%), $X^2=104$, df=1, $p<0.001$. The type of questions chosen is clearly very reactive to the method of selecting questions, that is whether they are chosen from a list or generated by the subjects.

The "*Confirmatory Bias*"

The third hypothesis was to see whether the questions selected conformed to the usual confirmatory bias. The prediction is that more biased extravert and open extravert questions would be asked to test the extravert hypothesis than the introvert hypothesis, and

more biased introvert and open introvert questions would be selected to test the introvert hypothesis than the extravert hypothesis.

Not one of these four main effects proved significant (all *F*s(1,36)<2.2, all *p*s>0.07, one tailed) but, as can be seen from Tables 7.1 to 7.4, they were all in the predicted direction. However, even when both categories of extravert questions and introvert questions respectively were added together to make two composite dependent variables, they were still nowhere near significant, both Fs(1,36)<1.9, both *p*s>0.1, one tailed. There were no significant effects for the neutral or irrelevant questions either (Tables 7.5 and 7.6).

Social Desirability

To check that there was consensus between the judges on their ratings of the "comfortableness" of the questions a form of Levi's (1974) "generalisability coefficient" was computed. This statistic gives an estimate of the amount of shared variance there would be in the scores of the questions used here and the scores expected if another group of eight rater-judges from the same population were to rank the same questions. The coefficient of 0.90 clearly demonstrates that the ratings are reliable.

The scores for the eight rater-judges were then averaged, and a one-way analysis of variance was computed to compare the social desirability of the six types of questions. This showed that there was a reliable difference between the six means, which are displayed in Figure 7.1, $F(5,42)=6.01$, $p<0.001$. Tukey's HSD test was employed to find which means were reliably different from the others. The biased introvert questions were found to be significantly less socially desirable than all of the other categories of questions (all p's<0.01), but there were no other differences (all p's>0.1).

A significant correlation between the social desirability of the questions and the number of subjects that chose to ask that question (Spearman's Rho=0.28, $p<0.05$, one tailed) demonstrates that social desirability is probably one of the influences on the questions asked.

Table 7.1. Mean Number of "*Biased Extravert*" questions selected (all *ns*=10)

Certainty of Hypothesis	Hypothesis: Extravert	Hypothesis: Introvert	
Hypothesis Only	1.4	1.8	1.6
Hypothesis Plus Expectation	2.4	1.5	2.0
	1.4	1.7	1.8

Table 7.2. Mean Number of "*Biased Introvert*" questions selected (all *ns*=10)

Certainty of Hypothesis	Hypothesis: Extravert	Hypothesis: Introvert	
Hypothesis Only	.7	.9	.8
Hypothesis Plus Expectation	.7	.8	.8
	.7	.9	.8

Table 7.3. Mean Number of "*Extravert*" questions selected (all *ns*=10)

Certainty of Hypothesis	Hypothesis: Extravert	Hypothesis: Introvert	
Hypothesis Only	1.5	.8	1.2
Hypothesis Plus Expectation	1.2	1.2	1.2
	1.4	1.0	1.2

Table 7.4. Mean Number of "*Introvert*" questions selected (all *ns*=10)

Certainty of Hypothesis	Hypothesis: Extravert	Hypothesis: Introvert	
Hypothesis Only	1.7	1.8	1.8
Hypothesis Plus Expectation	.8	1.8	1.3
	1.3	1.8	1.5

Table 7.5. Mean Number of "*Neutral*" questions selected (all *n*s=10)

Certainty of Hypothesis	Hypothesis: Extravert	Hypothesis: Introvert	
Hypothesis Only	1.6	2.3	2.0
Hypothesis Plus Expectation	1.9	1.4	1.7
	1.8	1.9	1.8

Table 7.6. Mean Number of "*Irrelevant*" questions selected (all *n*s=10)

Certainty of Hypothesis	Hypothesis: Extravert	Hypothesis: Introvert	
Hypothesis Only	1.1	.4	.8
Hypothesis Plus Expectation	1.0	1.3	1.2
	1.1	.9	1.0

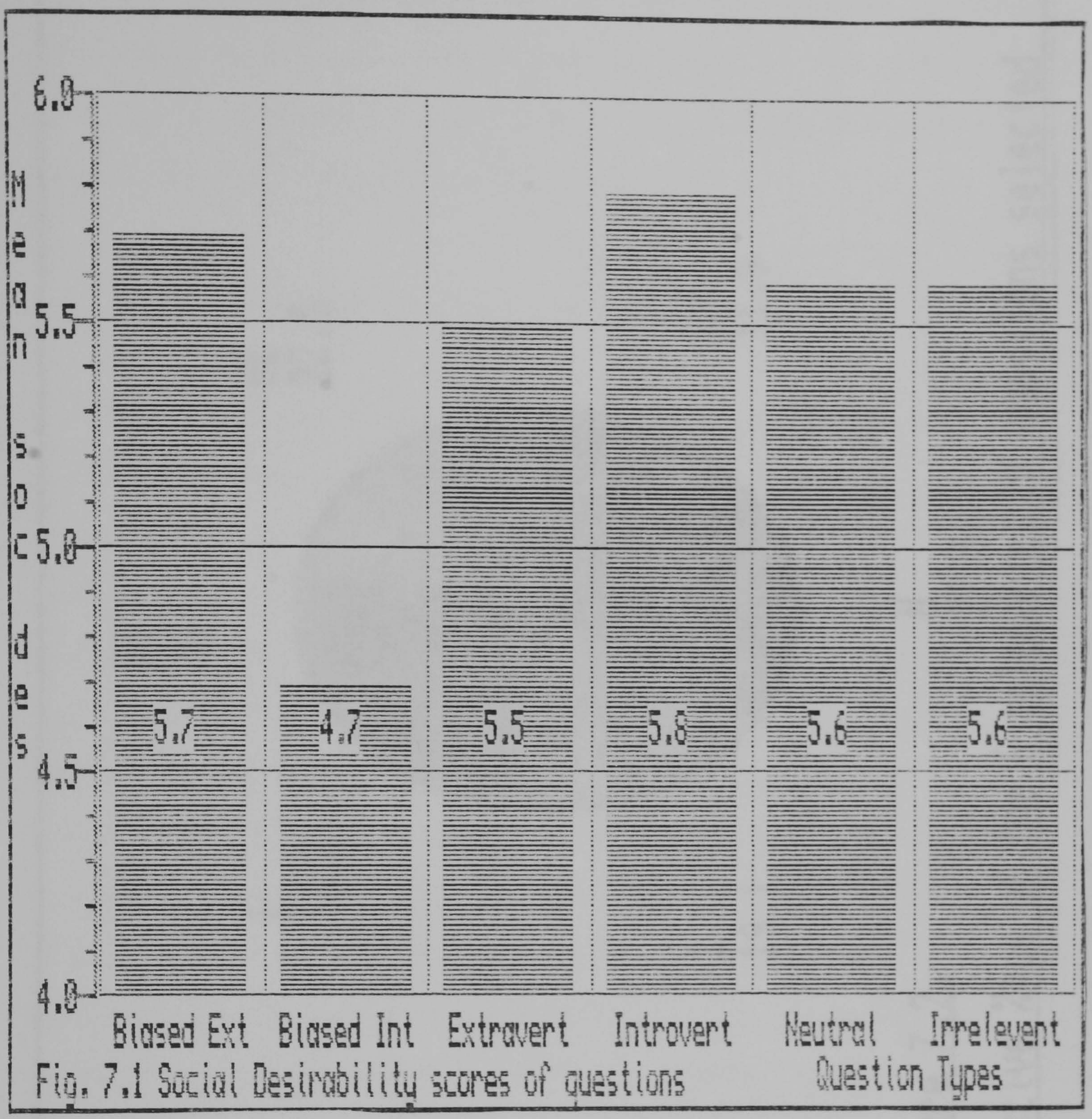

Fig. 7.1 Social Desirability scores of questions

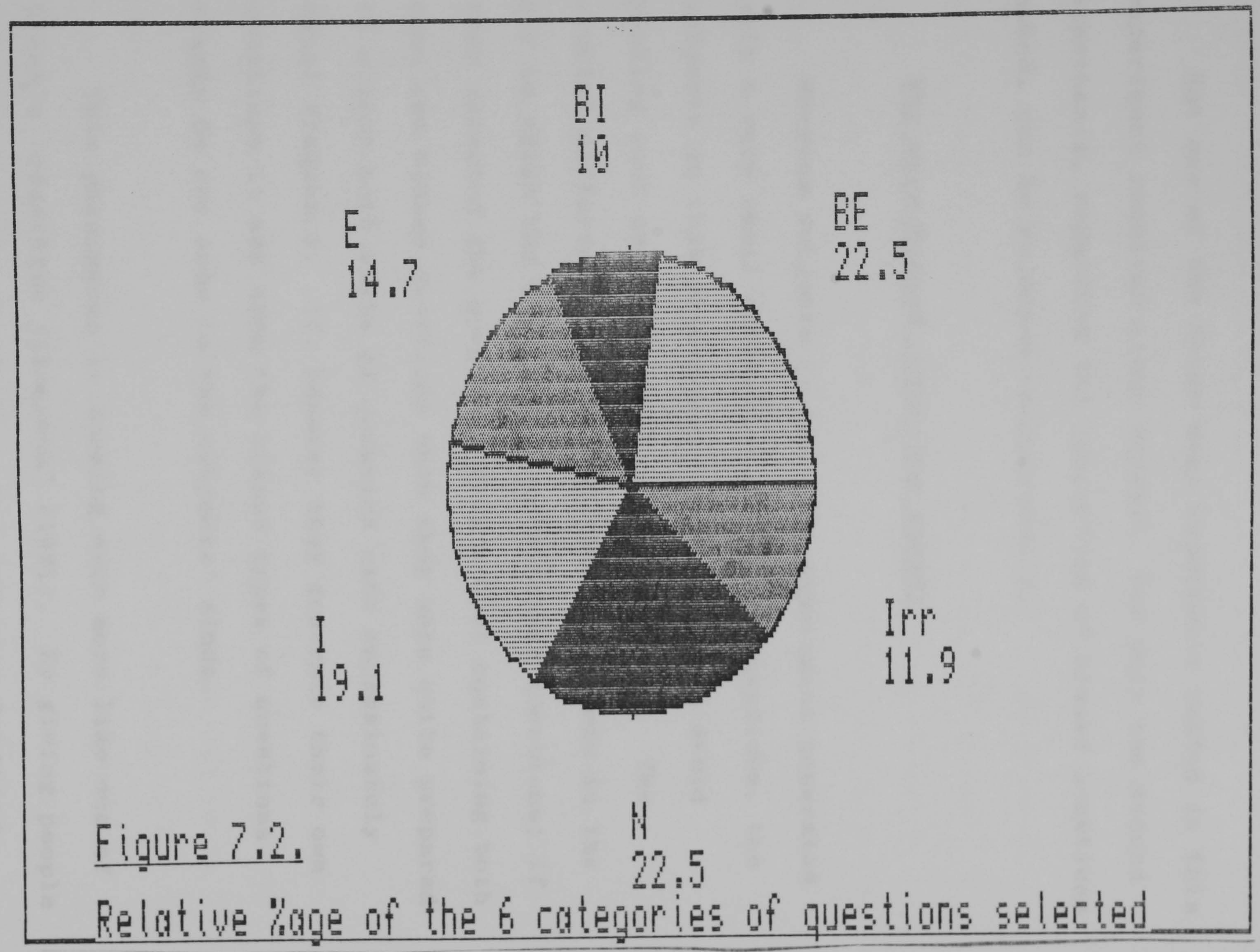

Figure 7.2.

Relative %age of the 6 categories of questions selected

Discussion

Not one of the three main hypotheses tested in this experiment received clear support, but only the second hypothesis, regarding the proportion of biased questions asked, can be rejected unequivocally.

Why were biased questions asked?

Whereas subjects in the last experiment generated only a very small proportion of biased questions, the subjects in this experiment did not seem to avoid picking such questions from the list at all. The crucial difference between the two experiments is the way in which the subjects selected their questions; if they selected the questions from a list containing both open and biased questions then they were quite prepared to select both sorts of question with approximately equal frequency. If, however they generate their own questions to ask then the biased types of questions simply do not come to the subjects' minds.

This phenomenon is looking even more like one of Cohen's "cognitive illusions" (1981). By giving people biased questions to ask, we are virtually "*putting words into their mouths*". Yet, although the concept of bias in questions is not readily appreciated by the subjects, it would seem as if they may become aware of it if they actually go ahead and ask questions. To support this, remember that some of the subjects who

were debriefed after asking their (biased) questions in the second experiment (ibid.) mentioned how the questions were leading or biased. These subjects spoke as if they were not originally aware of the bias in the questions, but became aware of the constraints as they asked those questions. Further evidence comes from an (albeit unsystematic) closer look at the interviews themselves. Often the interviewers would re-phrase a question during or after asking it, as if it was only by verbalising the question that the notion of bias became apparent. This is a typical transcript of a question from the first experiment (Chapter 4, ibid):-

Interviewer: *In what situations do you* (pause) *Well, do you sometimes wish you could be more outgoing?*

Target: *Mmm* (pause) *Yes, yes, sometimes I do I suppose,*

Interviewer: (interupts) *In what situations?*

The interviewer has, while in the process of asking the question "*In what situations do you wish you could be more outgoing?*" changed the question so that it is asked in two stages; the first stage checks that the premise is founded, and the second utterance on the part of the interviewer follows on from this and asks the remaining part of the question. This modified style of asking the question was not at all unusual in the interactions.

Thus, while it is easy to fool individuals into presuming that these types of question are alright to ask in some artificial situations, it is highly unlikely that biased questions would ever be unwittingly used in everyday interactions. It is worth noting at this point that (as argued in Chapters 3 and 9 ibid) even the asking of biased questions need not necessarily lead to biases in person perception. Indeed, the experiment in Chapter 5 found that they did not. More importantly though, if the choice of questions is that reactive to the exact method of choosing, then imagine how different the qustions asked during everyday social interactions are likely to be from the questions written down or selected to ask in the laboratory!

The Effect of Expectation

The first hypothesis, that the frequency of selection of the biased type of questions would increase with the certainty of the hypothesis, has received only moderate support. Although there was a significant effect for the biased extravert questions to be asked more often in the *expectancy plus extravert hypothesis* cell than in the corresponding *hypothesis-only* cell, the evidence from the biased introvert questions was in the opposite direction. In testing this "*effect of expectancy*" hypothesis the possibility of a *type 1* error cannot be ruled out, particularly given that only one of two *one-tailed* hypotheses was significant at the 5% level, an occurrence with a probability of almost

10% (or one minus 19/20 squared)[1].

Why no "*Confirmatory Bias*"?

The third hypothesis, searching for the all too familiar "confirmatory bias" did not receive any support from significant effects either. Although all of the relevant effects were in the predicted direction, they were very weak. One of the reasons for testing this hypothesis was simply as a check on the experiment. The "confirmatory bias" in question selection has now been replicated so many times and without failure (from the present experimenter's experience at least), that its absence here must lead us to question this experiment. Is something going wrong here?

The power of the present experiment certainly was not too low. Using an average of the sizes of effect found in previous experiments, (combined investigations 1 and 2 from Snyder & Swann, 1978, and Chapter 5, ibid.) the power of this experiment should have been in excess of 0.99 if alpha is set at 0.05 and with 40 subjects (Keppel, 1973, ch. 24).

So, were there any other important differences between this experiment and previous ones? One crucial difference may have been in the population from which the subjects were drawn. The experiments discussed so far on hypothesis testing in social interaction have, without exception, used exclusively (or in chapter 6

ibid., mostly) "normal" university undergraduates. Yet the subjects in this experiment were all Open University students. The post-experimental debriefing session revealed that many of these individuals had experience of interviewing in a personnel selection context, so they may have used more sound criteria in their selection of questions.

One study published recently is of direct relevance to this point. Sackett (1982) ran the now classic hypothesis testing paradigm using professional interviewers (Campus recruiters), and found that the hypothesis they were testing did not have any effect whatsoever on the types of questions asked. Experienced interviewers, he concludes, have well formed ideas about the types of questions that are useful in forming impressions of others in the interview situation, and are thus not likely to be affected to any great extent by the manipulation of the hypothesis. Other reasons could be that the interviewers are generally aware of the importance of an outgoing personality for many jobs, and thus automatically select questions to "confirm" this. It is interesting to note that in both Sackett's data and the data from this study, the ratio of biased extravert questions to biased introvert questions selected in all conditions is very high (2.4 : 1 and 2.3 : 1 respectively), compared to the results from chapter 5 (ibid.) or the combined investigations 1 and 2 from Snyder & Swann, (1978) (both 1.4 : 1).

An explanation for this high ratio of biased extravert to biased introvert questions may be in the social desirability of the questions. As the data from the rater-judges demonstrated, it is probably no coincidence that the biased introvert questions were selected least frequently of all the questions, even when including the "*irrelevant*" category, and they were also considered to be the questions that one would feel most embarrassed to ask. Perhaps the more mature and experienced subjects were also more aware of the importance of making the interviewees feel at ease, and of retaining one's own poise as a skilled interviewer.

If people with more experience of interviewing are less prone to the bias than others to whom the task is novel, then this might be an important but neglected consideration in the literature on shortcomings in human judgment. It is quite normal to use university undergraduates as subjects in the experiments then generalise the findings to some particular group such as interviewers or educators. If, however, skills learned by those experts make them impervious to certain biases, then this technique for research may well be inadequate.

The debriefing also revealed other sophisticated strategies that were being used by subjects in the selection of questions to ask. The results showed that the average number of "*irrelevant*" questions asked by each subject was one out of the eight. These questions,

for instance "*What do you think of Margaret Thatcher?*" or "*What is your opinion of marriage?*", were initially included as filler items, and their consistent selection was rather surprising. During debriefing the experimenter drew attention to these questions to see why they were selected. The subjects were readily able to defend their inclusion, usually by saying that either the question would make the target talk at length, thus revealing more about themselves, or by saying things like "*It would be a good question to start with, to get the interview flowing*".

These kinds of subtleties to do with the finer points and "rituals" of social interaction can easily be missed if an oversimplistic approach is taken to the social sciences (Harré, 1981). In our eagerness for the hard, mathematical rigor of the physical sciences we may miss the finer points that are so important in and characteristic of our everyday social lives. In the experimental social psychologist's endeavour to scrutinise social interactions in terms of information search and hypothesis testing, an important omission has cre pt in. In this paradigm social interaction is conceptualised as ahistorical in relation to individual lives, situation and cultural context, devoid of a dynamic development over time with spontaneity, affect and a multiplicity of goals for both participants. The present paradigm represents social interaction as a poor shadow of reality. If simplification is the goal of science, the hypothesis testing paradigms (and perhaps

others in social cognition) have overdosed!

Conclusions

1/ There is some weak evidence to show that biased questions may be asked more frequently when testing a hypothesis with a positive expectancy rather than simply a hypothesis that is as likely to be right as it is wrong.

2/ Contrary to the findings in the previous experiments, subjects chose biased questions with approximately the same frequency as other questions. It is concluded that the selection of questions is very reactive to the nature of the task, whether the questions are selected from a list or generated by the subject. Presumably the questions might be as different again if generated during social interactions.

3/ The "*confirmatory bias*" in the questions chosen was weak and not statistically significant. One possible reason for this may be that the subjects were not the normal university undergraduates but mature Open University students. Other experiments have also failed to replicate the *confirmatory bias* with people with interviewing experience. If the confirmatory bias only affects people without experience, this is yet another blow to Snyder's generalising of the laboratory effect to real situations.

4/ It was found that people felt less comfortable about asking the *Biased Introvert* questions than any other category of questions, and less comfortable questions are selected less frequently by the "interviewers". This can account for the reason why all of the experiments have found that biased extravert questions are selected more often than biased introvert questions.

1 This probability is arrived at by subtracting the probability of two non-significant results (19/20 squared) from 1, giving 0.0975. The assumption of complete independence is satisfied here, because the comparason is between different subjects, those in the extravert hypothesis condition and those in the introvert hypothesis condition.

A further possibility was looked into, that the effect could have partially been the function of an inappropriate parametric test on data that was, because of the very low means, not normally distributed or had unequal variances within different cells This is not the case, as was shown by a significant Mann-Whitney test on the same data, U=18.8, n's=10,10, p<0.01, one-tailed.

Chapter 8
Hypothesis testing
From Memory.

Introduction

The four experiments reported so far have been attempting to explore the way in which individuals test hypotheses about other people by asking them questions. As was mentioned in Chapter 2, this is not the only way in which individuals may go about searching for information to test hypotheses. This chapter departs from the theme of the previous chapters and looks at biases in the way individuals might test hypotheses but using information not gathered in social interaction. Another two possible mechanisms for testing hypotheses immediately spring to mind, testing hypotheses using information gathered by proxy or testing hypotheses using information already available in one's own memory.

In the first of these possibilities individuals may gather information from third parties. If, for example, I want to know whether another researcher is a good person to work with, I could ask questions of her co-workers. This would be an exactly parallel situation to the one investigated so far in that individuals may seek information in such a way as to collect an unrepresentative pool of knowledge about the target, and thus erroneously infer that the initial hypothesis was correct.

The second possibility is that we could already have a fair amount of information about the stimulus person stored in our memory, and simply try to remember information relevant to that hypothesis. Again, there is the possibility of biased recall leading to a predisposition to confirm hypotheses. For instance, say a collegue asked me "Do you think that the new Head of Department lacks confidence?" I might immediately have tryed to think of incidents in the past that confirmed the hypothesis. If, however, I had been asked to test the hypothesis that she was very sure of herself, I may have immediately tried to recall instances when she handled difficult situations with confidence, and again ended up concluding that the hypothesis was valid.

Snyder & Cantor's Experiments

The latter of these possible effectswas the topic of another of Snyder's long list of papers on hypothesis testing (Snyder & Cantor, 1979). Three experiments were reported in this paper, the first two being directly relevant here. These two experiments involved almost exactly the same methodology, the second one filling some methodological and conceptual faults uncovered while conducting the first. For current purposes, though, they can be described together.

Subjects read a lengthy passage about a ficticious woman called Jane. The passage depicted her behaving in a variety of ways; sometimes she behaved in a characteristically introverted manner (eg. being shy and timid at the supermarket) and sometimes she behaved in a way typical of an extravert (eg. not hesitating to talk to strangers while out jogging). In order to simulate the time factor involved in most person perception tasks, there was then a gap of two days before the second part of the experiment.

When they returned to the laboratory the subjects were told that Jane had applied for a job. In the *Introvert hypothesis* condition subjects were then told that the job that Jane had applied for was as a research librarian (a pretest had shown this to be perceived as a job highly suitable for an introvert). In the *Extravert hypothesis* condition the subjects were told that Jane had applied for a job as a real-estate salesperson (a job highly suitable for an extravert). Thumbnail sketches were provided to enhance the stereotypes of people suitable for these jobs.

They were then asked to try to remember and write down anything that would help them to use a six-point scale to indicate her suitability for that job. Finally, they were asked to rate her suitability for the <u>two</u> jobs, firstly the one that they were told she was applying for, then the other one.

There was also another, orthogonal, manipulation. Whereas some subjects were only told that Jane was simply applying for the job, others were told that she had been accepted for the job.

There was also a control condition in which subjects read the passage, but were not given any hypotheses to test or information about jobs. They were to provide the standard by which recall in the other conditions could be judged.

The results approximated to what Snyder & Cantor predicted. In both hypothesis conditions subjects reported more confirmatory incidents than hypothesis-disconfirming incidents. However, an (unpredicted) interaction between the hypothesis and the "*Applied-Taken*" factors showed that this effect was only reliable in the "*Applied For*" conditions, not the "*Job Taken*" condition. Comparisons with the control group showed that it was an increase in the reporting of hypothesis-consistent items that caused the "*confirm-atory bias*" effect, rather than a reduced reporting of hypothesis-disconfirming items. The authors tentatively explain the absence of confirmatory bias in the "*Job Taken*" condition as being caused by the subjects inferring that when she was offered the job that was already evidence enough of suitability for the post, and thus there was no need to seek further confirming evidence.

Subjects in all conditions then rated Jane as being better suited to the job that they had been told about initially rather than for the other job.

In interpreting these results Snyder & Cantor draw an analogy between the human mind and an archive library. When testing a hypothesis we sort through the contents of this library, and assemble all the data collected in this manner that we deem to be relevant before proceeding to analyse those data. They claim to have shown that there is a tendency to preferentially retrieve information that is consistent with the hypothesis under test.

They suggest several different mechanisms that may be responsible for this confirmatory bias. It may be because subjects believe that confirmatory information is more important than disconfirmatory information for a wide variety of tasks such as transmitting concepts (Hovland & Weiss, 1953), assessing covariation (Jenkins & Ward, 1965; Smedlund, 1963), evaluating propositions (Wason & Johnson-laird, 1972) or judging similarity (Tversky, 1977).

Their findings lead them to conclude that individuals are prone to exactly the same confirmatory bias when testing hypotheses using "historical" data as when they collect "fresh" data (ie. Snyder & Swann, 1978); hypotheses become self-confirming.

Using both the theoretical framework and empirical findings from this thesis so far, two major criticisms of this paper can be advanced. The first concerns the distinction between hypotheses and expectations, and the second deals with the artificiality of the experimental procedure.

Expectations and Hypotheses (revisited).

It has already been argued in Chapter 2 that hypotheses and expectations are conceptually distinct, and that this distinction is critical *in* evaluating the rationality of subjects in hypothesis testing experiments. To recapitulate, a hypothesis is merely a proposition about a feature of the world to be tested. An expectation, by contrast, is a prediction about the state of the world that gives information about likely states or outcomes. Whereas the particular hypothesis a subject is given to test should not affect the subject's final judgment, it is correct for an expectation to have such an effect. However, in exactly the same way as in Swann's PhD thesis (which it was argued in chapters 2 and 3, *ibid*, is a crucial link in Snyder & Swann's (1978) paper) Snyder & Cantor have completely confounded hypothesis testing and the effects of prior expectancies.

Subjects in this experiment have clearly been given a hypothesis to test, to see whether their knowledge of Jane makes her suitable for a particular hypothesis. But as well as being given this hypothesis, they were also told that Jane had either applied for or been accepted for a particular job. This, it will be argued, not only gave the subjects some plausible reason to test their hypotheses (which was presumably the authors' intentions), but it also gave the subjects some real information about Jane -- that she was probably an

extravert to apply for the job of real estate person, or that she was probably an introvert to apply for the job of research librarian. To deny that this inference could be legitimately made is paramount to saying that the types of jobs people apply for are totally unrelated to their personal dispositions. Even an individual with the gloomiest view of the accuracy of peoples' self-perceptions or of careers advisory services would, surely, not go that far! Furthermore, if we infer that a person is extravert because they apply for certain types of jobs, it follows logically that they will probably be less suitable for other jobs that require an introvert's disposition. So, instead of proving that individuals are displaying a shortcoming in their inferential powers by automatically confirming any hypothesis given to them, Snyder & Cantor's subjects may have ,instead, demonstrated that individuals are able to integrate expectancies into the hypotheses they test in a normatively appropriate manner according to Bayes Theorem. The very fact of applying for a job probably means that one stands a fairly high chance of being suitable for it. If a true self-confirming hypothesis is to be proven, then different hypotheses must not be confounded with different expectations or prior odds.

It is <u>not</u> being argued here that a confirmatory bias won't be found under these conditions, but simply that this crucial question is not answered by Snyder & Cantor's experiment. Indeed, the literature on long term memory structures suggests that several different

processes would account for a propensity to recall hypothesis-consistent information.

For instance, a network model of memory would predict that once a certain type of incident had been specified, (eg. extravert type incidents) similar incidents will be easier to remember since those nodes of memory will already be activated (see, for example, Hastie et al, 1980 for an extensive account of how a knowledge of the mechanisms of memory allows a greater insight into the processes of person perception). Thus the specification of an extravert hypothesis will stimulate more extravert memories in the brain, and make them more likely to be reported and subsequently used in the inferential process.

The artificiality of the experiment.

As well as demonstrating that an exact understanding of the information available to a subject be fully understood in hypothesis testing experiments, the thesis so far has also shown that some of the effects that work under very artificial conditions do not replicate when the tasks given to subjects are more realistic. The loss of the "*confirmatory bias*" in the sorts of questions that are asked when they are selected during rather than before an interaction was but one of many examples of this (Chapter 4, ibid). It would, therefore, seem prudent to make the experimental conditions as close as possible to the situation that one is attempting to generalise to.

The feature of Snyder & Cantor's experiment that is obviously most artificial is the presentation of information about Jane. While trying to simulate a store of knowledge that is built up over a long period of time and by many different types of experiences, Snyder & Cantor have simply presented subjects with a passage of information and asked them to read it. There is no evidence that recall from such a store of evidence is at all similar to a more naturally acquired set of information about a person. This point is similar to that made by Neisser when he states that "*The sentences and brief 'stories' that are popular in research laboratories today are an improvement on the nonsense syllable, but they are far from representative*

of what ordinary people remember and forget" (1982, p.11).

There is also strong empirical evidence that the knowledge built up from social interaction is qualitatively different from anything that can be simulated from other methods of data presentation. For instance, Gorman, Clover & Doherty (1978) found little evidence that similar processes are involved when interviewing real people and in "interviews" of "paper people".

It is hardly surprising , given the lengthy nature of the processes that occur in the formation of our memory about other people, that experimenters have tried to take short cuts in the testing of their theories. It would be prohibitive to suggest that all experiments like Snyder & Cantor's should be performed with the stimulus material being presented over a period of, say, several years. Fortunately, though, such processes are occurring all of the time outside of the laboratory in our everyday social lives. All that needs to be done is to bring these ready-made archives of information into the laboratory, to study the information-retrieval processes.

This is exactly what this present experiment is designed to do. The principle is quite simple -- subjects will be asked to test hypotheses about people that they have known for some time (*stimulus individuals*). Some will be asked to test the hypothesis that

the stimulus individual is an extravert, the other group will be asked to test the hypothesis that the stimulus individual is an introvert. After this hypothesis manipulation, the subjects will be asked to rate the stimulus individual on an extraversion-introversion scale. In addition, before the subjects make this rating of the stimulus individual, they will be asked to write down any particular memories that they have of the stimulus individual that they think relevant to testing the hypothesis. These memories will help to understand the recall and decision mechanisms that underlie the hypothesis testing processes. If, for instance, subjects report hypothesis-confirming and hypothesis-disconfirming evidence with equal frequency but then confirm their hypotheses, it could be inferred that the confirmatory bias occured at the information processing stage of the task. If, though, it is found that subjects preferentially report hypothesis-confirming data (as Snyder & Cantor found), then it can be inferred that the bias occurred at the data retrieval stage of the task.

The choice of stimulus individual was the next problem. It seemed reasonable to assume that if these biases do occur, then they may be more likely to occur when testing hypotheses about individuals that one is only moderately familiar with. If the stimulus individual person is too well known to the hypothesis testers, they may have a "pre-processed" reaction to the hypothesis and not need to go through the hypothesised

retrieval and processing stages. It was assumed that chairpersons of university departments might be suitable in this respect.

The following two hypotheses were made. Firstly, it was predicted that, in testing hypotheses about other people from memory, subjects would tend to report more hypothesis-consistent information than hypothesis-inconsistent information. Secondly, that they would rate the stimulus individual as being more extravert when testing the hypothesis that the stimulus individual is an extravert than when testing the hypothesis that the stimulus individual is an introvert (the introvert-extravert dimension was used again to make the results comparable with the previous experiments).

Method

Subjects

Fifty-two undergraduates from the University of Warwick volunteered to take part in this experiment. Thirty-two were from the department of Psychology and the remaining 20 from the School of Industrial and Business Studies (S.I.B.S.). There were 12 and 15 males in the two groups respectively.

Two rater-judges were employed, the experimenter and one other postgraduate psychology student.

Procedure

All of the experiment was conducted on a *Nascom 3* microcomputer, from the random allocation to groups through the giving of instructions to the administration of the *EPI*. Subjects received information from the Visual Display Unit (VDU) and responded using the keyboard or by writing down items on a pad of paper provided, as appropriate. The computer programme incorporated "*error trapping*" and other techniques to minimise the problems that a novice to computing might experience. The contact between the experimenter and the subjects before the experiment was thus limited to the brief introduction to the laboratory when the subject arrived thus minimising the possibility of *experimenter effects* (Rosenthal 1976).

Task

Firstly, subjects were told that the experiment was designed to investigate how people form impressions of others. Next they were told that the person they were going to be asked to think about was the chairperson of their department. The computer programme then randomly allocated them to either the *extravert hypothesis* or the *introvert hypothesis* condition and substituted in the name of their respective head of department to the instructional text.

In the extravert hypothesis condition the subjects were asked to decide how extraverted the stimulus individual was. They were then instructed to think in terms of concrete information about the way in which the stimulus individual actually thinks, feels and acts. Next they were presented with a description of a typical extravert to "*help them decide how much of an extravert*" the stimulus individual was (this was the same personality profile as used in chapters 4 to 7, *ibid*).

The subjects were then told to write down from six to twelve facts about the stimulus individual that would help them decide whether he was an extravert. They were given two examples to help them:

"(stimulus individual's name) *is usually quiet if he doesn't know the people he is with.*

and

"(stimulus individual's name) *is the sort of person who always talks to people on trains or in queues"*

When the subject had finished writing these items on the pad of paper provided, they rated the stimulus individual on ten 11-point (0-10) bipolar scales. These were the same scales as used in chapters 4 and 5, but scored in a different way because of their machine presentation. The scales were again summated to give a score between -50 and +50, higher scores indicating higher perceived extraversion.

Finally the subjects filled in a version of Eysenck's Personality Inventory (*EPI*) that asked them about the stimulus individual's extraversion. The questions were changed from the second to the third person, for example *"Does he like going out a lot?"*. Scores on this scale could vary between 0 and 24, higher scores again indication higher perceived extraversion. This scale was used again because the results from chapters 4 and 5 showed it to be more effective at accounting for variance than the summated scales (a more conventional way of measuring ratings of others).

The subjects were debriefed at length, and asked not to talk about it with other members of their department before all of the subjects had been run.

The task for the subjects in the introvert hypothesis condition was identical, except that the word *introvert* was substituted for the word *extravert* throughout the instructions, and a description of a typical introvert was given rather than a description of a typical extravert.

Rater-judges

Two judges independently read the examples written down by the subjects. Each statement was allocated to one of three previously agreed categories:-

1/ *Evidence of Extraversion*

2/ *Evidence of Introversion*

and

3/ *Evidence irrelevant to introversion or extraversion*

(neutral).

To remove the possibility of experimenter effects, the judges were kept naive as to the hypothesis conditions of the individual subjects.

The number of questions in each of these three categories plus the two measures of perceived extraversion made five dependent variables in total. All were analysed using a 2 (*introverted hypothesis vs extraverted hypothesis*) x 2 (*Dept. of Psychology Vs S.I.B.S.*) design.

Results

For the sake of simplicity the results will be divided up into two sections, firstly the incidents reported by the subjects and secondly the the subjects' final perceptions of the stimulus individuals on the *EPI* and summated scales.

The subjects' reported incidents

The first stage in the analysis of the types of incidents that the subjects reported was to check that there was sufficient consensus between the two judged. It was found that the judges agreed on the categorisation of 76% of the items. Cohen's Kappa was calculated to be 0.64, confirming that this was well above the level that would be expected from a chance-guessing model, $p<0.001$. The scores from the two judges were thus averaged to give one score for each subject on each of the three dependent measures, the numbers of extraverted, introverted and neutral incidents reported. Tables 8.1, 8.2 and 8.3 show the mean numbers of extraverted, introverted and neutral incidents in the four cells of the 2 (introvert vs extravert hypotheses) X 2 (Dept. of Psychology vs S.I.B.S.) breakdown.

These tables were analysed using three 2 X 2 ANOVAs. This is different from the analysis performed by Snyder & Cantor, who treated introverted and

extraverted items as a repeated measure in the analysis of variance. Their method, however, makes the assumption that the number of items reported in each category were statistically independent, and an exploratory analysis of the present data showed that they were clearly dependent.

More extravert statements were made in the extravert hypothesis condition (M=3.2) than in the introvert hypothesis condition (M=2.2), $F(1,47)^1=3.87$, $p=0.026$, one-tailed.

Similarly, more introverted incidents were reported in the introvert hypothesis condition (M=1.9) than in the extravert hypothesis condition (M=1.1), $F(1,47)=3.37$, $p=0.028$, one-tailed. There were no significant effects for stimulus individuals or stimulus individual by hypothesis for either of the introvert or the extravert incidents.

There was no significant difference between hypothesis conditions with respect to the neutral incidents. The only significant effect was between stimulus individuals, but this is difficult to interpret because it confounds department of undergraduates with stimulus individual. Anyhow, the effect is of no relevance to the purposes of this experiment.

The Subjects' final perceptions of the stimulus individual.

Tables 8.4 and 8.5 give the *EPI* and summated scales scores respectively of the subjects' final perceptions of the stimulus individuals. There were significant main effects for the hypothesis manipulation using both measures. The stimulus individuals were seen as more extraverted in the extravert hypothesis condition on the *EPI* (M=12.4) and the summated scales (M=14.1) than in the introverted hypothesis condition, *EPI* (M=10.1) and Summated scales (M=7.8), *EPI* $F(1,48)=3.41$, $p=0.035$, and Summated scales $F(1,48)=4.46$, $p=0.02$, both one-tailed.

The *EPI* did display a significant difference between stimulus persons, showing that the head of the Psychology department was perceived to be more extraverted than the head of S.I.B.S.. There were no other significant main effects or interactions.

The proportion of variance accounted for by the hypothesis manipulation was also calculated using the procedure outlined by Vaughan and Corballis (1969). The ω^2 statistic was found to be 0.06 and 0.04 for the summated scales and the *EPI* respectively, meaning that the hypothesis manipulation only accounted for 6% or 4% of the total variance in the subjects' perception of the stimulus individual. When compared to the variance found to be attributable to the targets' stable

disposition in Chapter 4 (between 26% and 42%), this effect is almost an order of magnitude smaller. So while the effect is statistically reliable, it may still not be an important process in person perception.

1. The incidents reported by one subject were lost, thus the loss of one degree of freedom in these analyses.

Discussion

The two hypotheses tested in this experiment received unequivocal support. When asked to test a hypothesis about an acquaintance, subjects reported more hypothesis-consistent data than hypothesis-inconsistent data. Following this, subjects were clearly biased by the hypothesis that they were testing when making judgements about the veracity of the hypotheses; subjects reported that the stimulus individual was more extravert when testing an extravert hypothesis rather than an introvert hypothesis. This confirmatory bias was apparent using both of the measures employed, the *EPI* and the Summated scales; not only did the subjects rate the stimulus individuals differently on bipolar adjective scales, but they also answered the *EPI* questions differently for them on a wide variety of topics (unfortunately, because the *EPI* scale was scored automatically by the computer, scores to individual questions are not available and thus it is not possible to determine exactly which particular questions were primarily responsible for this effect).

At last a true self-confirming hypothesis seems to have been demonstrated, and without having to resort to obviously artificial experimental methods. Individuals have come to perceive others, previously known to them, to be consistent with a hypothesis that they were given to test.

Four points of interest come out of these findings; the possible mechanisms for the effect: a comparison with experimental findings from other research into the effects of hypothesis testing and expectations on recall from memory: an important statistical point to arise from the methodology used: and finally the relevence of the effect demonstrated here to everyday person perception. These points will be discussed in that order.

The mechanisms for the effect.

Before considering how and why the hypothesis manipulation caused a bias in memory recall or inferences from that recall, a reference to the accounts given by the subjects will be useful. Each subject was carefully debriefed, both before and after the experimental hypothesis was explained to them. One common sentiment expressed was that subjects thought themselves to be guessing, and they went to some length to point out that they considered some of their responses to be almost totally lacking in evidence to support them. In particular some subjects singled out questions from the *EPI* and said that from their limited knowledge of the stimulus individuals (mainly from watching them lecture), they had almost no way of knowing how to answer items such as "*Would he do almost anything for a dare?*" or "*does he suddenly feel shy when he wants to talk to an attractive stranger?*". If there had been a *don't know* response category it would

almost certainly have been used frequently.

An examination of the incidents that the subjects had written down also showed that they often had difficulty in generating the six to twelve items requested. From behind the two-way mirror in the laboratory it could be seen that subjects were spending a long time on, and agonising over, this particular stage of the experiment. Talking with the subjects confirmed this. When asked about some of the incidents that subjects had reported, it was clear that they were almost pure guesses, often derived very obviously from the personality profiles that were presented. For instance, when one subject was asked why he had reported a particular item "*I would expect him to enjoy a quiet drink/meal with close friends rather than a boisterous party*", the subject responded "*I hardly know him -- I thought that was the sort of thing you wanted*". This almost total and deliberate fabrication of evidence was only apparent for a very small minority of the subjects, but it does highlight the problem of demand characteristics. This effect may be similar to the one noticed by Rush, Thomas & Lord (1977). They found that the validity of questionnaire responses is threatened when subjects have little real information to go on and use their "implicit theories" to provide plausible responses.

In addition, it suggests a point that may be of theoretical interest. The subjects seemed to have even

more difficulty in reporting the specific evidence upon which their evidence was based than in making the personality judgments, suggesting that Snyder & Cantor's "*archival library*" (p. 331) model of memory may be a poor characterisation of the actual processes involved. Do we really select evidence, assimilate it, then make an inference in that order?

There is some evidence that deciding first then rationalising why we decided that way is a better representation of human epistemic processes. Zajonc (1980) has argued that simple affective preferences can be made before any cognitive processing has occurred. While Zajonc's cognitive verses affective distinction may not be the same as Snyder & Cantor's "remember then decide" model, it does call into question any over-simplistic models of cognitive processes. Most models of memory suggest that concepts are stored hierarchically with specific behaviours nearer the bottom and generalised traits further up. If this is the case then individuals may be able to access the traits directly, without having to re-process the behaviours each time. The way the task was organised for the subjects may have forced them to use their inference in a manner different from their normal procedures. Other considerations of models of human memory may also be of assistance in trying to understand the hypothesis testing process.

One such model is the network model. Human memory is hypothesised to consist of a network of inter-

connected nodes. The nodes represent concepts and the connections represent relationships between concepts. When an individual is exposed to a stimulus, concept nodes directly associated with that stimulus are activated. When nodes are activated, the excitation spreads to other nodes closely connected by the relational pathways. When enough excitation accumulates at these secondary nodes, they too become activated (see Chapter 3 in Wyer & Carlston (1979), for a detailed description of this model). One of the benefits of this model is that it leads to a number of specific hypotheses that can be tested. Many of these relate to "priming effects", the effects of the inferences made from an earlier stimulus on later information processing. This is caused by residual excitation making some nodes much more predisposed to becoming excited again.

Of specific interest here is Wyer & Carlston's (1979) discussion of what happens when individuals are asked to decide whether a particular entity is a member of a given category. This is similar to the task facing subjects here as they attempt to decide whether their stimulus individual is a member of a category (ie. the category of extraverts or the category of introverts). By firstly activating one category (ie. the extravert or introvert hypothesis and associated personality profile) the concept nodes associated with that category will be activated. Some of this excitation may not have decayed completely when the entity is considered soon after-wards. Any nodes that are excited by the entity and

already posess some residual excitation are much more likely to attain their activity threshold, and thus become excited again. In other words, any concepts that may be linked to both the entity and the category are much more likely to be activated than entities linked to one but not the other. The first of these categories corresponds to hypothesis-confirming information, the second to hypothesis-disconfirming information.

One fundamental difference between this mechanism and the one proposed by Snyder & Cantor is the level of control that the subject has over his or her own epistemic processes. Snyder & Cantor talk about these processes as being predominantly under the control of the individual, who occasionally, for whatever reason, display\ shortcomings in those processes. Wyer & Carls on's network model, by contrast, characterises the individual as being forever and inextricably handicapped by the very structure of his or her mind. If one long-term goal of this research is to know how to train people to think in a more rational manner then it is essential to understand the level at which the bias displayed in this experiment operates.

A comparison with other person memory experiments.

There is a considerable socal psychological literature on the way in which individuals encode and retrieve information about others. This literature is full of apparently inconsistent and opposing findings. Before attempting to compare the present results with other findings, that literature will be reviewed, albeit very briefly.

The typical experiment on encoding and recall biases is concerned with the effects of expectations of groups or individuals. Typically the information available to the subjects about the stimulus material is manipulated (eg by the use of stereotypes) so different groups of subjects have different expectations. To attempt to distinguish between biases arising during encoding from those occurring at the retrieval stage, the time at which the expectation is created is often also manipulated. If the expectation is created *before* the stimulus material is presented, then encoding and retrieval processes are both liable to be affected by it, but if the expectation is created *between* the presentation and recall stages of the experiment, then any biases found can only have occurred at the retrieval stage.

Some studies have found evidence of bias at the recall stage. For instance Snyder & Uranowitz (1978) presented biographical material about a woman, then

later told the subjects that she was involved in either a heterosexual or homosexual relationship. On recall, the woman's past was "reconstructed" in line with the stereotypes of heterosexuals and homosexuals. Other research has only found such biases when the expectation was created before the presentation of the material, and thus concluded that expectations can only affect encoding, not recall (eg Rothbart, Evans & Fullero, 1979). Even the direction of the effect has not been entirely predictable -- Hastie & Kumar (1979) found that expectation-inconsistent information was more likely to be remembered than information consistent with the information. It seems as if the differences in the procedures employed by different experimenters could potentially be the causes of these inconsistencies. For instance, in some experiments the subjects were explicitly told that it was a memory task, whilst in others the subjects were prevented from attempting to memorise the stimulus material. The complexity of the material varied greatly between the experiments, as did the relative frequency of the consistent and inconsistent items. In some experiments there was a gap of several days between the presentation and the recall stages, in others they followed each other almost immediately. Sometimes the stimulus material was presented at the level of traits, in others the "raw" behaviours were presented. Some experiments used a free recall procedure, others used specific or multi-choice questions. To date, the relative contributions of these factors can only be guessed at (Berman, Read & Kenny, 1983).

Given all of these complexities, it is impossible to say whether the results of this experiment are compatible with those from other studies. In addition to all of the differences between studies listed above, there is a further important way in which the experiment presented here is different from all of the others. In chapters 1 to 3 (*ibid*) it was argued that hypotheses and expectations were conceptually different, and the conclusions from a study of one cannot be generalised to the study of others. The same holds true for hypothesis testing from memory. The experiment presented here was concerned with the effects of hypotheses and not expectations, and therefore the mechanisms of information encoding, storage and retrival being studied could be very different from studies where expectancies are manipulated. It should come as no surprise, however, that the two types of experiment are often talked about interchangably (eg. see Berman, Read & Kenny, 1983).

Another difference between the present experiment and those reported in the "person memory" literature involves the task of the subjects. In most of the person-memory experiments subjects were simply told to recall as much of the material that was presented as possible (a notable exception to this is Hamilton & Gifford, 1976). By contrast, the subjects in the experiment presented here not only had to remember as many details as possible, they were also required to assimilate the material and make inferences from it.

This may again require a different process; remembering an impression may be very different from remembering the reasons for that impression.

While there have been no other experiments in the literature explicitly about hypothesis testing from memory, many of the questions that have been asked about the effects of expectations on memory could also profitably be asked of hypothesis testing. For instance, what happens if the hypothesis is given to a subject before the encoding of the memories? In some ways this would be a similar situation to the "hypothesis testing in social interaction" one, but without the opportunity to intervene and manipulate the target directly, only to interpret and encode the target's behaviour selectively. The cross-fertilisation could occur in the other direction too; most of the person-memory experiments would be much more realistic if they used a methodology more similar to the one in this experiment, with real people as the stimulus individuals rather than "*paper people*" or lists of traits and behaviours.

A statistical consideration.

Before going on to discuss the generality of this finding, a weakness of the present design and other designs used in research on person perception will be discussed.

One of the fundamental points about inferential statistics is that they allow the generalisation of the effects from the specific subjects used in the experiment to the population from which they were selected. In other words, the confirmatory bias demonstrated in this experiment does not just apply to the 52 subjects who happen to have been selected to take part, but it is possible to say that the effect would be the same with any similarly selected group of university undergraduates.

Things are often more complicated than this in person perception because experimental trials often involve more than one subject (eg a perceiver and a target) rather than just one. This is best handled by treating both of the categories of subjects as a single random factor, as was done in the experiments in chapters 4 and 5 (ibid). This involves having a different pair of subjects for each trial, and, statistically speaking, is no different from the normal "single subject" design. Another situation often used in social psychological experiments is to have just one subject per trial, but to use deception to create the illusion

of another subject (as was done in chapters 6 and 7, ibid). Again, this is straightforward, and needs no special statistical consideration; the actual and imaginary subject are both different for each trial, the imaginary subject only existing in the mind of the real subject.

Problems arise, however, when the stimulus individual in these types of experiments is specified, as in Snyder & Cantor's experiment. While the subjects are a true random factor, the stimulus individual (Jane) is the same for all trials. This means that, while any significant effects can be generalised to all perceivers, they cannot be generalised to any other stimulus individual but the imaginary Jane (not even to other "paper people"). Whilst we can predict that, when testing the real estate salesperson hypothesis any subject would tend to remember more extraverted things about Jane, it is technically incorrect to infer anything about the situation when they test a hypothesis about any person except Jane. If we want to make this inference, it is not a statistical inference but instead we must fall back on our knowledge as psychologists and argue that the effect is not specific to any one stimulus individual, but will be the same for all stimulus individuals. Clearly in some situations we would not be prepared to generalise from one individual to the population, in other situations we might.

The experiment reported here is slightly more complex in as much as there are two stimulus individuals. Now it would be possible to treat both the perceivers and the stimulus individuals as random factors, instead of treating the stimulus as a fixed factor as was done here.

The new error term for the *F* ratios would become the mean square of the interaction between the hypothesis factor and the stimulus individual factor. The problem with this is that it would cause a devastating loss of power. With only one degree of freedom for the error term, an effect would have to be enormous to be statisticallly significant. To overcome this problem the number of stimulus individuals would have to be increased, preferably to about the same as the number of subjects.

Alternatively, the data could have been considered as two separate experiments. This, however, would not overcome the problem of generalising to the population of stimulus individuals, it would only have allowed inferences about the way Psychology and S.I.B.S. students test hypotheses about their respective heads of departments.

A much neater solution to this problem would have been to have a different stimulus individual for each subject. When designing the experiment this solution was rejected for two reasons. Firstly, it would

probably have reduced the power of the experiment markedly because differences in "actual" extraversion of the stimulus individuals could not have been disentangled from the confirmatory effect, so would have had to be included in the error term. Secondly, it would have been difficult to instruct the subjects to choose stimulus individuals that were equally well known. Perhaps, though, with hindsight, this alternative solution would have been preferable. The subjects could have been asked to test hypotheses about, for instance, the head teachers at their school before they came to university.

But let it not be forgotten that, while these reservations are being expressed, the methodology used here is still almost certainly better than in virtually all of the other person-memory experiments. What it comes down to is where we as psychologists feel most comfortable making generalisations. Is it better to generalise from a written description to real people (ie. Snyder & Candor, 1979), or from descriptions like "*Robert was rated highly on the trait friendly*" to real people (ie. Berman, Read & Kenny, 1983), or from a series of trait descriptions and sentences to real people (ie. Hastie & Kumar, 1979) or from two real people to other real people? I think that most psychologists would agree that people are more similar to other people than imaginary profiles are to people.

Hypothesis testing in the real world.

As well as the theoretical interest in a bias affecting the inferences we make about other people, the effect demonstrated here should have important implications for the real world. In this section the power and pervasiveness of the phenomenon will be discussed. Firstly, how powerful an effect is this confirmatory bias on testing hypotheses from memory?

Power

The ω^2 statistic showed that only four to six percent of the variance in perceived extraversion was attributable to the manipulation of the hypothesis (using the *EPI* and summated scales respectively). It is not possible to compare the present results directly with the data from the experiments in previous chapters, but as a yardstick remember the variance attributable to "genuine" differences in the extraversion of the targets when the *EPI* was used to divide them into two groups in chapter 4. In that experiment the target's extraversion accounted for 41% of variance in the interviewer's ratings using the *EPI*, or 25% using the summated scales. Thus the bias looks as if it may be of only theoretical interest being "drowned out" by other more important factors. But would the power of the effect be larger or smaller in more naturalistic situations?

Theories have been put forward to argue this both ways. On the one hand, it could be said that the subjects were trying to be "*on their best behaviour*" in the laboratory, when they knew that they were being studied. Prehaps in the real world hypotheses are tested in a more haphazard and less structured way, an environment where biases may flourish better than in the laboratory.

On the other hand it could be argued that the laboratory situation has been set up with the express purpose of finding a confirmatory bias. The hypothesis was presented in such a way, with the personality profile and no mention of the alternative hypothesis, that should maximise any such bias. In addition the stimulus individuals were chosen to be of just the right familiarity with the subjects to create the right ambiguity for the effect to work.

Clearly, more experiments are required to find the environmental facilitators of the effect, and to determine under what conditions it might be strong.

Pervasiveness

In the present study the experimental instructions lead the subjects through a hypothesis testing task. If the findings are to be of applied importance, then it is also important to know when individuals actually test hypotheses in this way. Perhaps the most straight-

forward occasion when this may occur is when someone asks our opinion of a third party. If I were to ask you "*Do you think that John is an extravert?*" or "*Do you think that Jane would make a good research librarian?*" then I am explicitly asking you to test a hypothesis.

The other situation in which hypotheses may be tested is when we set ourselves hypotheses to test. For whatever reason, I may suddenly wonder whether a new friend is, in fact, rather shy, or I may wonder whether the fact that an old friend has not been in touch for a long time means that she does not like me any more.

Which of these two situations is similar to the present experiment? The answer is probably that we do not know, but it could be argued that the experiment is different in important respects from both of these situations. Consider the latter example. When people set themselves hypotheses to test, how are they framed? Do people test hypotheses in the same way as in this experiment, or do they generate several alternative hypotheses and test between them? Perhaps when we ask ourselves questions they are in the form "*Why hasn't Mary been in touch with me recently? Is it because she doesn't like me any more* <u>*or*</u> *because she is involved with a new boyfriend* <u>*or*</u> *is it because it is my turn to write to her?*". If hypothesis testing tasks were set in this way then the testing process may be different. There is little evidence in the literature on the way people set themselves hypotheses to test. In Chapter 6 (ibid),

though, it will be remembered that subjects, in testing hypotheses with their own questions, overwhelmingly asked questions that seemed to be testing between two hypotheses (eg. "Do you prefer reading books or talking to people?") or that would provide a response useful for comparing between two hypotheses (eg. "What do you do in your spare time?"). If the questions that people spontaneously generate to ask other people are of the same sort that they generate to ask of themselves, then the process demonstrated in this experiment may be very different to the way in which people spontaneously test hypotheses themselves. Furthermore some evidence was also found in chapter 6 that presenting both sides of a hypothesis made subjects even more likely to search equally for both confirmatory and disconfirmatory evidence.

And what of the other example of when we might test hypotheses, when someone asks us specific questions about another person? On the surface it seems much more likely that the processes involved there mirror the present experiment more closely. But asking questions of others is not a simple "*question - process - respond*" chain, particularly when the question is fairly complex (like here). It seems much more likely that these types of situation will involve dialogue, and (as was demonstrated in Chapter 4), dialogue can have a marked effect on information search and hypothesis testing, diverting individuals away from a confirmatory strategy.

The outcome of this discussion leads to the conclusion that we know little about the pervasiveness of the confirmatory bias. In fact it is possible that the hypothesis testing process that subjects went through in this experiment is unlike any process that they normally use, and therefore may tell us little about everyday person perception. Before firm conclusions can be made about when exactly the effect found in this experiment will bias people's perceptions of others it will be necessary to find out more about the way in which hypotheses are set and framed in their everyday lives.

Conclusions

1/ In this experiment individuals were asked to test hypotheses about individuals they already knew. It was found that they would tend to search for evidence supportive of the hypotheses and then conclude that their hypotheses had been confirmed. This seems to be a genuine bias in person perception, made all the more impressive by the use of real stimulus individuals and without the confounding of hypotheses and expectations that marred Snyder & Cantor's experiment.

2/ Several possible criticism of the present experiment were considered. It seemed as if the demand characteristics placed on the subjects may have been partly responsible for the subject's responses. The

effect may also have been enhanced by the way in which the hypotheses were presented to the subjects. More research is needed to explore these factors.

3/ The results of this experiment are compared to other findings in the person-memory literature. The present findings are comparable with the dominant models of memory, but there are so many inconsistencies in the findings from person-memory experiments that it is difficult to draw specific conclusions. Besides, this experiment was investigating the effects of hypotheses on recall, other experiments have been primarily concerned with expectations.

4/ The various ways in which experiments involving pairs of subjects (eg perceivers and stimulus individuals) are handled statistically in Social Psychology were discussed. It was concluded that the best way to handle this is to treat both as random factors, and that this is most easily achieved by having a different stimulus individual for each perceiver. The technique of using just one (usually fictitious) stimulus person is the least satisfactory solution. The method used in this experiment of having a small number of real stimulus individuals, while not being perfect, probably provides a good balance between generalisability and power.

Further consideration of the relevence of these experimental findings, along with a discussion of wider issues such as rationality and normatively correct ways of testing hypotheses will follow in the next chapter.

Table 8.1. Mean Number of *Extravert* incidents reported

		Hypothesis		
		Extravert	Introvert	
S.I.B.S.		M = 3.4 SD= 2.1 n = 10	M = 2.1 SD= 1.4 n = 9	M = 2.8
Stimulus Individual	Psychology	M = 3.1 SD= 2.0 n = 16	M = 2.3 SD= 1.9 n = 16	M = 2.7
		M = 3.2	M = 2.2	

Table 8.2. Mean Number of *Introvert* incidents reported

		Hypothesis		
		Extravert	Introvert	
S.I.B.S.		M = 0.6 SD= 0.6 n = 10	M = 1.6 SD= 1.7 n = 9	M = 1.1
Stimulus Individual	Psychology	M = 1.4 SD= 1.7 n = 16	M = 2.1 SD= 1.7 n = 16	M = 1.8
		M = 1.1	M = 1.9	

Table 8.3. Mean Number of *Neutral* incidents reported

		Hypothesis: Extravert	Hypothesis: Introvert	
S.I.B.S.		M = 3.2 SD= 1.7 n = 10	M = 3.1 SD= 1.7 n = 9	M = 3.2
Stimulus Individual	Psychology	M = 1.3 SD= 0.8 n = 16	M = 1.7 SD= 1.1 n = 16	M = 1.5
		M = 2.1	M = 2.2	

Table 8.4. *EPI* ratings of the two stimulus individuals

Stimulus Individual	Hypothesis: Extravert	Hypothesis: Introvert	
S.I.B.S.	M = 12 SD= 2.9 n = 10	M = 7.4 SD= 3.4 n = 10	M = 9.6
Psychology	M = 12.8 SD= 5.2 n = 16	M = 12.7 SD= 2.9 n = 16	M =12.8
	M = 12.4	M = 10.1	

Table 8.5. *Summated Scales* ratings of the two stimulus individuals

Stimulus Individual	Hypothesis: Extravert	Hypothesis: Introvert	
S.I.B.S.	M = 18.5 SD= 6.6 n = 10	M = 7.7 SD= 11.4 n = 10	M =13.1
Psychology	M = 9.6 SD= 11.0 n = 16	M = 7.9 SD= 10.3 n = 16	M = 8.8
	M = 14.1	M = 7.8	

Chapter 9
Theoretical Issues from Other research and Conclusions.

Introduction

Taken together, the picture of the testing of hypotheses painted by the five experiments presented in this thesis is very different to the one portrayed in the existing literature. In this chapter other literatures from Psychology that have a direct bearing on hypothesis testing in the social world will be considered.

This will start with a consideration of works that explore the issue of diagnosticity and use a Bayesian framework to evaluate subjects' performance. Then the attribution research paradigm will be compared and contrasted to the hypothesis testing one. Following this the results obtained in this thesis will be considered in the light of the literature on rationality, and the issues of biases, errors and optimal performance will be considered. Then other experiments are suggested that would contribute to a continuing advancement of knowledge in the fields of hypothesis testing and of self-fulfilling prophecies and conclusions will be drawn, not only about the narrower issues considered in this thesis, but also the wider issues concerning the way in which research is conducted

and reported that led to such an erroneous understanding of hypothesis testing processes. It will be suggested that this paradigm may not be alone in having painted such too bleak a picture of human rationality, and that the methods of enquiry employed in much social cognition research needs to be reviewed.

Bayesian Analyses and Diagnosticity in Hypothesis Testing.

There is another set of experiments in the social cognition literature that, instead of being only concerned with a possible confirmation bias, is also concerned with the efficiency of information search. In searching for information with which to test hypotheses, some information will be better than others. That is, some information will allow the hypothesis tester to become much more knowledgable about the veracity of the hypothesis, whereas other information might be of little or no use in that respect. Take, for example, the following two questions "*Do you prefer reading books or talking to people?*" and "*Do you prefer reading books or having a quiet night in watching television?*". In structure they are identical, giving the target a choice between two alternatives. In the information they are likely to yield pertinent to testing a hypothesis about extraversion, though, they are very different. The first of the questions is good because one would expect different answers from an extravert and an introvert, and therefore a preference for reading books can be taken as

evidence for introversion and a preference for meeting people can be taken as a evidence of extraversion. Judged by the same criteria the second is poor, because there is probably no difference between introverts and extraverts in their likely response to the question, and thus the inquirer is left no wiser about the target's personality after asking the question. In this respect the first question can be called diagnostic, and the second non-diagnostic.

Several experiments have looked explicitly at the question of diagnosticity in information search behaviour. One set of experiments that explicitly tested

whether subjects were primarily concerned with confirming their hypotheses (as Snyder & Swann, 1978, found) or with obtaining diagnostic information was published in 1982 by Trope & Bassok.

In a series of experiments Trope & Bassok's subjects were instructed to test the hypothesis that another person possessed certain personality traits, whether they were intuitive thinkers or analytic thinkers. Subjects were told that it is possible to tell which of these a person is from their handwriting. They were given a booklet that contained the results of a survey by graphologists of handwriting among these two personality types. The data were broken down by eight different features of handwriting, such as variability in writing angle and margin. The proportion of individuals in the two groups with each particular feature was given

in a graphical form. For instance, a bar graph may have indicated that 20% of analytical thinkers showed variability in margin, compared to 78% of the intuitive thinkers.

The data about each of the eight features of handwriting were presented in a seperate bar graph. Under each graph there was a rating scale on which the subjects indicated how interested they were in each bit of information, on an 11-point scale from "not at all interested" (0) to "very interested" (10).

Two factors were manipulated orthogonally, the diagnosticity of the criteria and the probability of the feature being present under the hypothesis. If subjects follow a confirmatory hypothesis testing strategy, they would rate as interesting those features of handwriting that are most likely to be present under the hypothesis being present. If, however, subjects follow a diagnostic strategy, they will prefer those criteria where there is a large difference between the probability of the feature being present under the hypothesis compared to the alternative hypothesis (Trope & Bassok's definition of diagnosticity is, in fact, slightly more complicated than this, but the exact definition and formulae are not important to the present discussion).

The results of the experiments demonstrate that, while there was a significant confirmation strategy, the effect was exceedingly small compared to the subjects'

preference for diagnostic information.

In a follow-up to this study, Trope & Bassok (1983) used a question-selecting paradigm to demonstrate that diagnosticity is still the most important feature of a question, and again the preference for confirmatory information was found to be weak by comparason. In addition, they explored how the exact way in which a hypothesis is set affects the types of questions that are asked. They point out that in Snyder's hypothesis testing experiments, the hypotheses were always phrased in terms of what an extreme extravert or an extreme introvert is like. It is then ambiguous whether testing, say, the extravert hypothesis the task is to judge whether the target is an extreme extravert or not, or whether he or she is simply more of an extravert than an introvert.

When Trope & Bassok set the task in terms of discriminating between intermediate positions rather than extremes, the confirmation bias disappeared. Not only was the diagnosticity of the questions a more important factor under all conditions, but subjects even showed that they were able to appreciate that different questions were more or less suitable depending on the hypothesis "boundary", that is whether they were testing between *extreme extraverts* and *not extreme extraverts* or between *extraverts* and *introverts*. An additional, welcome feature of their experiments was that they also varied the trait that they were using, and showed that

the effect replicated not only for the well-worn introversion-extraversion dimension but also on a polite-impolite scale.

Bayes Theorem

Some other experiments have also drawn upon aspects of Bayes theorem as a normatively correct framework for hypothesis testing. To understand these experiments, though, it is neccesary to understand the theorem in more detail. It will, therefore, be described briefly before the contributions of these other experiments are evaluated.

Bayes theorem is so named after the Reverend Thomas Bayes who first derived the formula and published it in 1763. The formula shows how probabilities should be updated given new information. In its simplest formulation it can be expressed thus:

$$\underbrace{\frac{P(H/D)}{P(\bar{H}/D)}}_{(A)} = \underbrace{\frac{P(D/H)}{P(D/\bar{H})}}_{(B)} \times \underbrace{\frac{P(H)}{P(\bar{H})}}_{(C)}$$

where H is the hypothesis, and D is a new datum gained from an observation. $\bar{H}$ represents the the alternative hypothesis. P(H) means the probability of the hypothesis being true, and P(H/D) means the probability of H being true given that D is true.

The three components of the formula can be read as follows. C is the *prior odds*, the probability of the initial hypothesis being true divided by the probability of that hypothesis being false. B is the *likelihood*

ratio, the amount by which the the prior odds is updated given the new datum. A represents the *posterior odds*, the new, updated, odds of the hypothesis being true. In person perception experiments, the prior odds can often correspond to expectancies that a perceiver has about a target. The likelihood ratio would then correspond to the amount of updating of the hypothesis after new information is learned by, for instance, observing or asking questions of the target. When the preceiver is asked for his or her final evaluation of the target after the new information, this corresponds to the posterior odds.

An example will show how the Bayesian formula should be used to modify probabilities. Imagine a woman moves into the house next to mine, and I want to know what her politics are, whether she votes Conservative or not. Before I collect any more information I may already be able to make an informed guess from what I know about her. If I knew that she was a self-employed businesswoman I might guess that the chances are, say, 60% (or 0.6) that she votes Conservative (ie, $P(H)=0.6$, so $P(\bar{H})=0.4$). The prior odds of her voting Conservative are thus $0.6/0.4=1.5$.

The next day I may be out in my garden, and have the opportunity to talk to her. It is soon after the Libyan raid, so I say over the fence *"What do you think of the bombing of Libya then?"*, knowing that 25% of Conservative voters were favourably disposed to it

compared to only 10% of the rest of the population. Suppose that she responds by showing her disapproval of the raid. The likelihood ratio is now the chance of this datum (disagreeing with the raid) occuring under the hypothesis (75%) compared to the likelihood of it occurring under the alternative hypothesis (90%), 0.75/0.9=0.83. Multiplying this by the prior odds, we get 1.5 X 0.83 = 1.25, the posterior odds. This means that my new best estimate is that the chance of her being a Conservative has dropped to 56% (0.56/0.44= 1.25). This makes intuitive sense, she is less likely than before to be a Conservative, but not by all that much since Conservative voters were not that dissimilar to the rest of the population in respect to their view of the raid on Libya. If, instead, I asked her whether she thought that Mrs Thatcher was doing a good job (knowing that 90% of Conservative voters think she is compared to 10% of the rest, say) and the neighbour said no, then the chance of her being a Conservative voter would now drop from 60% to 14%, a much bigger fall.

This represents the diagnosticity issue put forward by the Trope & Bassok nicely. The first of my questions was low on diagnosticity since my perception of the target was hardly different after collecting the information than before it. The second question was highly diagnostic because I learned a lot about the target, as displayed by the large change between the prior and posterior odds. In this respect the highly diagnostic questions are better than those with low

diagnosticity, and Trope & Bassok demonstrated that individuals prefer those questions likely to produce highly diagnostic answers.

The results from the hypothesis testing in social interaction experiments reported in this thesis can also be interpreted as supporting the view that individuals are good intuitive Bayesians.

It will be recalled that in the first two experiments the interviewers were quite accurate in their assessments of the targets' "actual" extraversion as measured by the w^2 statistic, regardless of the hypothesis that they were testing. If they had been collecting information in an entirely undiagnostic manner, however, they would not have been able to assess it so accurately.

The avoidance of "biased" questions and the preference for "neutral" questions found in chapter 6 is also suggestive of a correct Bayesian approach. "Neutral" questions, those that give the target free reign to answer in an manner indicative of their true personality, are probably the most diagnostic, whereas questions that constrain all targets to respond in a similar manner, like the "biased" questions, are the least diagnostic.

But the debriefing of subjects in Chapter 7 revealed that individuals may be much more subtle in

their approach than Trope & Bassok's experiments allowed for. The "Irrelevant" category of questions would presumably have been classified as completely undiagnostic by Trope & Bassok (1983), but this would have completely ignored the richness of social interaction and the sophistication with which individuals might manipulate it. The strategies revealed in the choosing of questions, such as asking questions to "*bring people out of their shell*" or "*put them on the spot*" demonstrate that the optimal strategy in the social world is not necessarily the one that conforms to simple mathematical rules. Trope & Bassok's modeling of social interaction processes (1982, 1983) is even more removed from reality than Snyder & Swann's (1978) in this respect.

In contrast to Trope & Bassok's and these experiments, others have demonstrated severe shortcomings in the layperson's use of information to update expectancies. For instance, Doherty, Mynatt, Tweney & Schiavo (1979) and Beyth-Marom and Fischhoff (in press) have found that, under some circumstances, subjects collect and utilise the wrong information to perform a proper Bayesian analysis, an effect they call *pseudodiagnosticity*. More specifically, they found that, when subjects had to request information to assertain the likelihood ratio, they would first collect p(D/H) but then neglect to request $P(D/\bar{H})$. In the example given above, this would be equivalent to asking the question about the raid on Lybia, knowing the proportion of

Conservatives who supported the action but being ignorant of the probability that a non-Conservative would support the action. Since the likelihood ratio cannot be calculated without $P(D/\bar{H})$, the information that was collected is totally non-diagnostic, or useless.

It is also possible to see the under-utilisation of base-rate information (eg. Kahneman & Tversky, 1973) as another deviation from normative, Bayesian principles; base rates are, after all, simply prior odds.

The research comparing human performance to the principles of Bayes theorem seems to provide mixed conclusions. In some experiments the subjects conform well to the correct procedures as determined by the formula (eg. Trope & Bassok, 1982, 1983), in other experiments the subjects are shown up as almost incapable of testing a hypothesis (Doherty et al, 1979). Why this discrepancy?

While no conclusive answers to this question are available, a tentative suggestion can be proposed. It is apparent that sometimes only minor changes in the wording of questions can have significant effects on the subjects' judgments of the relevance of information. The way in which performance can be raised and lowered so simply with logically equivalent tasks is remenicent of another set of rules sometimes obeyed and sometimes contravened by individuals -- the *if ... then* propos-

ition. In that case and the present, a normative model for thinking seems to be on the threshold of lay understanding. When presented in concrete terms (eg. Johnson-Laird, Legrenzi & Sonino Legrenzi's, (1972) deductive reasoning task with first and second class stamps or finding out about other people in social interaction as in this thesis) then subjects perform well. By contrast, when the problems are set in terms not familiar to the subjects (eg. Wason & Johnson-Laird's (1972) card-turning task or Doherty et al's (1979) characteristics of archaeological finds from hypothetical islands) subjects perform badly as if they completely misunderstand the logic of the task.

As well as generating new experiments to test different aspects of hypothesis testing, Bayes theorem can also be used to clarify our understanding of Snyder & Swann's experiments. Two aspects of this will be discussed next.

What is a *Confirmation Bias* in information search?

Snyder & Swann's principal conclusion from their experiments on hypothesis testing was that subjects ask questions that search for evidence supportive of the hypothesis under scrutiny. Fischhoff & Beyth-Marom (1983), however, point out that from a Bayesian perspective, it is far from clear exactly how this is possible by simply asking questions.

According to Snyder & Swann, 21 of the 26 questions on their list were of the "biased" sort, which already assume that the target is either an introvert or an extravert. Since the subjects had to choose 12 questions, they were forced to ask a majority of these unsuitable questions. Now consider what happens when a subject asks one of Snyder & Swann's *biased extravert* questions. The assumption is that when a question like "*What would you do if you wanted to liven things up at a party*" are asked, most individuals, regardless of their actual extraversion or introversion, will give similar answers, talking about behaviours typical of an extravert. If the answers of subjects are the same whether the hypothesis is true or false, then that is the same as saying that $P(D/H)$ and $P(D/\bar{H})$ have the same value, making the likelihood ratio equal to 1.

This is not to say that it would be impossible for the interviewers to conclude that their hypotheses had

been confirmed; rather, (as Fischhoff & Beyth-Marom argue) it means that any bias that is present is occurring at the stage of data interpretation rather than data collection. Subjects may, for instance, ask questions where a particular response is probable whether the hypothesis is true or false. If subjects then ignore or under-utilise the denominator in the likelihood ratio, $P(D/\bar{H})$ (as was found in the *pseudodiagnosticity* studies) they may thus falsely infer that the hypothesis has been supported. Fischhoff & Beyth-Marom's point is that this is an error in information processing, not in information search.

Since Snyder & Swann's interviewers were not asked to process the responses to their questions (see Chapter 3, ibid) it is wrong of Snyder & Swann to conclude that a confirmation bias had occurred. Furthermore, if they had asked their interviewers to interpret the tagets' responses, the evidence from Chapter 5 (ibid) demonstrates that they probably would not have made any such error.

Klayman & Ha (1986) also argue that the term "*confirmation bias*" has been misused in referring to phenomena like Snyder & Swann's question selection task. They argue that it is better described as a "positive test" strategy or heuristic, which is not necessarily dysfunctional as the term "*bias*" implies. They claim that in many situations it is a good heuristic to look for positive instances, but in the question-asking

situation it can lead to one important oversight -- that many questions that search for positive instances may be nondiagnostic, particularly when forced to choose from a list of biased questions.

Hypotheses and Expectations

Another issue that can be seen very clearly with Bayes' theorem is the distinction between hypotheses and expectations, and the normatively correct utilisation of expectancies in judgments.

As described earlier, an expectation corresponds exactly with the prior odds in the formula. It is clear from the formula that the prior odds should have an effect on the posterior odds, except when the likelihood ratio is either zero or infinity, (that is, when a datum is conclusive proof of the truth or falsehood of the hypothesis, which rarely occurs in the ambiguous social world). So, the mere fact that in Snyder, Tanke & Berscheid's experiment (1977) or Swann's Ph.D. thesis (1978) the perceivers still held their expectations after the interactions is not in itself evidence of a bias in information processing. Whereas Swann refers to the use of initial impressions in the final judgment rather derogatively as "*parroting back*", it is in fact entirely consistent with the rules of logical information processing as decreed by Bayes' theorem -- not to do so would clearly be wrong.

What would constitute a bias is over- or under-utilisation of the prior odds. Unfortunately, though, it is rarely possible to put exact values on any of the terms in the formula when studying social interaction, because of the lack of objectivity in interpreting the exact meaning of social behaviour. The advantage of using other materials, such as the category membership data expressed in percentages in Doherty et al's (1979) or Beyth-Marom & Fischhoff's (1982) studies is that it allows the correct answer to be specified exactly. In studies of truly social phenomena, the only time a deviation from the theorem can be proven is if subjects modify their beliefs or expectations in the wrong direction. For instance, Bayes' theorem dictates that any difference in prior odds between two perceivers should be moderated by the presentation of similar data. The easiest way, therefore, to demonstrate the existance of a bias is to show that different interpretations of similar data cause subjects' posterior odds to be more divergent than their prior odds. In Chapter one it was demonstrated that this has yet to be done.

The hypothesis under scrutiny, by contrast, should not affect the posterior odds (Burchell, 1984). The easiest way do demonstrate this is by seeing what happens when the hypothesis and the alternative hypothesis are interchanged. This is equivalent to testing the introvert hypothesis rather than the extravert hypothesis. The effect on the formula is to

swap H for $\bar{H}$, which simply turns the whole hypothesis upside-down, but the formula retains exactly the same form, notably:

$$\frac{P(\bar{H}/D)}{P(H/D)} = \frac{P(D/\bar{H})}{P(D/H)} \times \frac{P(\bar{H})}{P(H)}$$

It can also be seen from the formula how hypotheses can be tested even if their probability is very low -- for instance a routine cancer checkup on a healthy patient could be seen as a hypothesis test with a very low prior odds.

It can be seen from these discussions that Bayes' theorem provides a clear framework for research into hypothesis testing. It is not being argued that great new insights can be gained from the theorem that cannot be argued from "common sense" (as was done in Chapters one to three), but it is being proposed that Bayes' theorem provides a convenient and concise way of formulating research into information search and hypothesis testing issues, and of interpreting the results obtained. This analysis suggests that if Snyder and his collegues had used a Bayesian framework to conceptualise the logic behind their experiment they would not have made so many errors in designing their experiments or in interpretating of their results.

Attribution Theory and the Testing of Hypotheses

Since attribution theory is the dominant framework for research in person perception, it is perhaps surprising that it has all but been ignored in the information search and hypothesis testing literature. Why should this be?

In this section differences in emphasis and focus between attribution theory and hypothesis testing will be discussed. Not only will this highlight differences between the two approaches but, more interestingly, it will allow some of the criticisms that have been levelled at attribution research to be addressed to the hypothesis testing literature.

The similarities and differences between making attributions and testing Hypotheses.

Attribution research is primarily interested in the process by which behaviour is explained by the lay-person. The raw data is behaviour, and the explanations are usually in terms of personal dispositions (cf "*from acts to dispositions...*", Jones & Davis,1965). The hypothesis testing literature, is also concerned with the link between behaviour and dispositions. It takes as its starting point a task being given to the subject, who has to decide which behaviours should be recalled or sought to test the hypothesis as set (in terms of traits). Thus while attribution research is concerned

with <u>why</u> things happened, and hypothesis testing is concerned with the veracity of traits, which in turn can be seen as generalised causes of behaviours. Given this similarity between the two fields, it is all the more surprising that there has been so little cross-fertilisation of findings and ideas.

From a functionalist perspective there is also a lot of similarity between the two fields. The primary aim of subjects when testing hypotheses and making attributions is seen as prediction and control in line with the "man as scientist" role. Furthermore, the maintainance of self-esteem has been studied in both attributional (eg Greenwald, 1980) and hypothesis testing (Swann & Hill, 1982b) terms, another possible goal of the two processes.

Perhaps one of the biggest contrasts between the fields of attribution theory and the "judgement under uncertainty" literature (including hypothesis testing) was the competence attributed to the individual. Early attribution theory in particular (eg. Kelley, 1967) represented people as using elaborate statistical models to arrive at their judgements. By contrast in the classic research on judgement under uncertainty the subjects frequently show themselves as incapable of making even the simplest of decisions (Tversky & Kahneman, 1974; Nisbett & Ross, 1980). Reading the two different accounts one might have even found it hard to believe the two camps of research were using the same

species in their experiments!

This difference was probably caused by the early attribution researchers asking the question "*What do people do*" and the judgement under uncertainty researchers asking "*what do people do <u>wrong</u>?*". More recently, however, there has been a convergence of the two fields as attribution theorists have become interested in biases in the attribution process which has lead to some very productive research (steming from Ross, 1977 and others).

<u>Criticisms of Attribution Theory applied to hypothesis testing.</u>

While it beyond the scope of this thesis to present a comprehensive critical evaluation of attribution theory, there are several lessons that can be learned from past research within the attribution paradigm that can be constructively applied to the hypothesis testing paradigm.

One of the major areas of uncertainty in the models of human thought proposed by attribution theories is <u>when</u> exactly individuals frame their thought processes in line with the classic attributional models (Eiser, 1983). There is now an abundance of evidence from experiments that, when given the appropriate information to process, subjects integrate the information in a way that is by and large consistent with

these models. For instance, McArthur (1972) gave subjects distinctiveness, consensus and consistency information and asked them to explain an act in terms of the person, entity, circumstance or some combination of these three, which they, with some notable exceptions, managed to do.

But, to demonstrate that individuals <u>can</u> use information in this way is completely different from demonstrating that people <u>do</u> use information in this way in their everyday lives. One of the problems that researchers have consistently found when trying to use the findings of attribution research in naturalistic or applied settings is that individuals only sometimes seem to spontaneously express their interpretation of events in the categories deemed suitable by attribution researchers. For instance Antaki (1982) found that ordinary language expla nations often bore little resemblance to the dependent measures in experiments such as McArthur's. Other research has found that individuals only give "*Why*" answers to questions in certain particular circumstances. For instance, Wong & Weiner (1981) studied spontaneous information search and explanations in a variety of contexts. They concluded that, while people do engage in attributional search without prompting, this tendency is greater in some situations than others, namely unexpected events or failures. It could even be argued in this case that, by asking subjects to find *causes* of events, Wong & Weiner had already suggested that "attributional"

information was to be sought rather than any other type of information (Eiser, 1983).

The parallels of these criticisms with the hypothesis testing paradigm are clear. When do individuals spontaneously use social interactions to test hypotheses? When they do test hypotheses in social interactions, how do they frame those hypotheses? There has been hardly any attempt to address either of these questions to date. The only exception is a brief discussion on the pervasiveness of hypothesis testing in an article by Snyder & Gangestad (1981). They point out that the notion of man as hypothesis tester has been prevalent in many areas of psychology such as cognition and perception. Furthermore, Snyder & Gangestad draw upon the arguments of philosophers of science who have argued strongly that all thinking must be guided and preceeded by hypotheses, that "blind induction" is not possible.

Even if this is true, the questions of what sort of hypotheses and how those hypotheses are selected to test remain. It has already been argued in Chapter 8 that hypothesis testing from memory may not necessarily be framed in the same way as Snyder & Cantor (1979) assumed that it was. The arguments are, if anything, even more forceful when hypothesis testing in social interaction is considered.

Throughout all of his work on hypothesis testing Snyder assumes one model of the philosophy of science, and infers that lay epistomology works on the same principles. His model is similar (though not identical) to Popper's falsificationist view. Snyder assumes that individuals should hold one hypothesis and test it by seeking both evidence that is supportive of the hypothesis and evidence that is inconsistent with the hypothesis. Presumably, were this hypothesis to be rejected, the individual can then find another hypothesis to test, and so on. Thus unlike Popper, Snyder does not see falsification as the main aim of hypothesis testing. In fact, Snyder never makes it explicit what strategy a subject should use, except to point out that verificationist, falsificationist and "equal opportunities" strategies could all lead to errors.

Snyder never makes it clear why individuals might use this "single hypothesis" approach in their quest for knowledge. In tapping lay epistemological processes, Popper's prescription for science is probably a poor starting place. In considering the falsification strategy, Popper was not describing the "natural" way to go about testing hypotheses, but rather advocating a method that scientists should use, and saying that the progress of science has been hampered in the past by the failure on the part of scientists to clearly articulate their theories before testing them. There is clearly no justification for Snyder's implicit assumption that what

scientists <u>should</u> do is what lay people <u>do</u> do.

This is an important point. If subjects are being put through a wholly artificial procedure in following the experimental task, then the findings of those experiments, while possibly of theoretical interest, have no direct implications for the real world, making all of the applications of Snyder & Swann's research by themselves and others completely unfounded and misleading.

How else, then, might individuals frame hypotheses? There is no direct evidence that allows an answer to this question, but it is quite plausible that there are completely different ways in which individuals might use hypotheses to structure their thinking. One possibility is that individuals start their thinking process by generating multiple hypotheses, and then collecting evidence that eliminates these hypotheses successively, and perhaps suggesting others, until they are left with only one. This process was proposed as a model of the "efficient" sciences based on historical observation rather than on logical grounds by Platt in 1964, who called it "*strong inference*". Why, then, did Snyder use this particular model of hypothesis testing without questioning its suitability as a model of processes in the social world? The answer to this question can perhaps be found in the early cognitive experiments on hypothesis testing. It will be argued, however, that hyopthesis testing in the physical world and hypothesis

testing in the social world are sufficiently different as to make any direct inferences from one process to the other of dubious credibility.

This "multiple hypothesis" approach as a model of human thought is inconsistent with the findings of early cognitive experiments on hypothesis testing. For instance, the hypothesis testing task in which subjects had to discover the rule governing the series of numbers starting with 2, 4, 6... led Wason (1960) to conclude that the failure to be able to eliminate old hypotheses and move on to new ones was a severe shortcoming in human cognition.

In a later study Mynatt, Doherty & Tweney (1977) did find some evidence that at least some subjects could use a multiple hypothesis strategy as suggested by Platt. In this experiment subjects were put into a computer-simulated research environment where they were instructed to discover the rules governing that environment. Mynatt et al deduced from their subjects' responses during the task that only about half of the subjects seemed to formulate alternative hypotheses. Furthermore, the way in which the hypothesis or hypotheses were framed was important in determining the success of the subjects on the task; subjects who started with hypotheses that at least mentioned the important variables were much more likely to be successful in their task than the others whose initial hypothesis was totally incorrect. Thus these two

studies both concluded that human hypothesis testing was characterised by a failure to reject old hypotheses and to move onto new ones; instead subjects kept trying to find new evidence for their existing hypotheses.

There may, however, be an important difference between the types of hypotheses in these experiments dealing with the asocial material and the types of hypotheses tested about other people. In trying to discover the laws governing a physical environment there are a very large number of possible hypotheses that have to be processed. The problem is not so much in testing the accuracy of a hypothesis once specified, but rather in the generating of the correct hypothesis. It is thus a creative task, to generate better hypotheses than the ones rejected. It was not unusual for subjects in Wason's task to know that their current hypothesis was wrong, but be unable to think of a better one to replace it. Similarly, in scientific research often the faults, shortcomings and inaccuracies of a current theory are known long before a better one is proposed to dislodge it.

The nature of hypothesis testing in Snyder's social interaction experiments is very different from this. The number of hypotheses is very limited -- either the target is or is not an extravert. The task involves evaluating a given specific hypothesis, not going through a series of hypothesis generation and rejection stages to eventually arrive at the correct answer.

This all points to the fact that Snyder's characterisation of the way in which individuals test hypotheses about each other may be totally different from any actual naturally occuring hypothesis testing processes. Individuals may, for example frame their hypotheses in terms of competing hypotheses, such as "*I wonder whether he is an extravert or an introvert?*". While this is only a speculation, it is no less plausible than Snyder's speculation which, it is argued, not only draws invalid parallels between social and asocial hypothesis testing, but also misrepresents the possible ways in which hypotheses may be tested in the physical world. Furthermore, there is evidence in an unpublished paper that when hypotheses are presented in this "*equal opportunity*" way the "confirmation bias" is eliminated (Cooper, 1982).

A further consideration about the way in which individuals test hypotheses in the social world concerns the genesis of those hypotheses. In Snyder's discussions of the relevance of hypothesis testing to everyday social interactions he gives examples of situations in which individuals might test hypotheses (these are described in more detail in Chapter 2, ibid). Most of his examples take the form of a perceiver receiving some new expectation upon which he or she forms an impression of another individual, and then proceeds to test this expectation as a hypothesis. For instance I may hear that a new lecturer is an extravert, so test for

extraversion when I next meet her. This type of process has one main weakness. It again confounds hypotheses and expectations, and ignores the fact that people may test hypotheses with low prior odds. This is particularly odd since Snyder demonstrated that not only are people capable of testing hypotheses with low prior odds, but also that they still look for confirmation under those circumstances (Snyder & Swann, 1978, investigation 3).

A second situation in which hypothesis testing may take place is when, for some reason, there is a new need to know certain information. For instance, if I were to consider going on holiday with an old friend, I may think that the last person I would like to go on holiday with would be someone who is not good at meeting new people. I may therefore decide that I have to test the hypothesis that this new friend is good at meeting new people. In this situation, the circumstance has brought about the need to test a hypothesis.

But both of these situations are ones that attribution theory (eg Kelley, 1972) suggests would elicit very different processes to Snyder's question asking. Is it possible that there is some chance of a reconciliation here?

Kelley postulates that different attributional processes will occur depending upon the amount of information about the target available to the perceiver.

If the perceiver has a lot of information readily available, he or she will use Kelley's ANOVA model. By contrast, if perceivers are working from limited data, they will use causal schemata. These are theories that individuals hold about the likely causes of particular events. I may, for instance, have several causal schemata to account for students failing exams, such as lack of effort, general lack of ability, poor exam technique, etc. Instead of a full inductivist search for all of the possible causes of the event when a student fails an exam using Kelley's consensus, distinctiveness and consistency information, I may simply test within this limited range of likely causes. Shaklee & Fischhoff (1982) looked at the situation in which subjects had to test between several different potential causes of an event. For instance, they were told "*Tim advertised for a room-mate to share his apartment*" and were given three possible causes, financial, loneliness or fear of crime. Shaklee & Fischhoff found, when given the opportunity to find out more information to determine the cause(s), that subjects would spend much more time looking for further details about a cause that they thought more probable. For instance, in the example above, if they were told that Tim's scholarship had been cut, they would be more likely to ask for further informtion about his rent but if they were told that his long-term relationship had recently broken up they would search for more information about his possible loneliness.

Shaklee & Fischhoff also found that the subjects search patterns were of the "*limited truncated*" type. Thus, instead of simultaneously exploring all possible causes they would concentrate on them one at a time, and finish when they had satisfied themselves that they had found a sufficient cause, even if they had not fully explored all the other possible causes. This process, they argued, could cause individuals to erroneously conclude the expected causes were the only causes of an act, even when there may be other necessary or contributory causes as well.

The strengths of this type of hypothesis testing paradigm over Snyder's is that it integrates well into the rest of the person perception literature, whereas it is difficult to fit Snyder's account into either the rest of the person perception literature or the unconstrained natural environment.

Attribution Theory and "Mindlessness".

Another criticism of attribution theory focuses on the way in which it assumes that individuals invest a lot of cognitive effort into the processing of information. Langer (1978) argued that everyday actions are often not as well thought out and planned as attribution theory assumes.

To support this assertion, Langer conducted several experiments to show that people were not processing the information content of messages so much as simply reacting to their structure. For instance, she demonstrated that important changes to the content of a message (eg a legitimate or illegitimate request to jump a queue) did not effect compliance with the request, but the addition of completely redundant information (eg. explaining that one needs to make copies, when in a queue for a photocopier) could affect compliance. Langer argues that individuals simply do not have enough processing power to utilise the models of thought proposed by attribution theory. Instead, they probably follow "scripts" (Abelson, 1976) for most of their well-learned social environments. While there have been criticisms of Langer's empirical methods and the generality of her conclusions, (Harris & Harvey, 1981) there is still a lot of validity in her warning not to presume too much mental activity in people's everyday lives.

The same criticisms are equally appropriate to experiments on hypothesis testing. According to Snyder's accounts of the hypothesis testing process several stages of carefully planned thought are involved. Firstly individuals have to formulate a hypothesis, then they have to generate questions to test that hypothesis, before even interacting with the target. It is perhaps more plausible that the stages of finding out specific things about other people are rather more "mindless" than this.

Like the previous point about the way in which hypotheses are framed, it is very difficult to discover the cognitive processes that occur in everyday interaction, yet social cognition researchers seem to make such assumptions implicitly all the time. Because they are assumptions they are often above the level of testable propositions, and thus can be very enduring. The danger is that the processes being studied in social cognition bear little relationship to the processes that occur in our everyday social lives.

Hopefully the findings of the experiments in this thesis have shown how unrealistic and misleading some of Snyder & Swann's (1978) assumptions were, but even in the experiments presented here the hypothesis testing process may be much more explicit than the everyday occurrences that they are trying to emulate.

The Rationality Issue.

So far in this thesis some of the main points about the rationality of the subjects in experiments and of possible strategies for hypothesis testing have been deliberately side-stepped. Far from being the simple issue that it first appears to be, such words as "irrational" and "biased" have become the topic of debates, and often heated debates at that (eg. see Cohen, 1981, including peer commentaries). Several key arguments from the literature on human rationality and Snyder's own arguments concerning biases and errors will both be reviewed before attempting to draw conclusions about "rational strategies" for hypothesis testing.

Abelson (1976) argues that social psychology's cautious use of the term "irrational" can be traced back to the studies of authoritarianism in the 1950's. To call a system of values, beliefs and goals irrational can take the "objectivity" away from Social Psychology, and open them to the accusation of political bias.

The use of the term "rational" in that sense is very different, though, from the normal use of the term in social cognition. In that example, to say that someone is irrational is to mean that their picture of reality is distorted. While it may often be tempting to label people with deviant political or religous views as irrational, there we are referring to the assumptions upon which their reasoning is based, not the actual

reasoning itself. The type of rationality that is relevant to social cognition is concerned with processes of thought, not with beliefs.

Baron (1985) makes a similar point, but put rather differently, by saying that rationality can be considered hierarchically. At the highest level is the choice of a rational *life plan*, taking into account personal and moral interests. Below this level there are medium level policies that individuals choose, and at the lowest level are individual decisions. The upper levels of the hierarchy determine the goals and utilities of the lower levels, but otherwise the processes involved are very different. According to this model it is only the lower levels of the hierarchy that are the concern of this thesis.

It has been pointed out that psychology's model of man varies greatly over time with the amount of rationality ascribed to it. Sometimes man is portrayed as being capable of even very complex logical and statistical deductions and inferences (eg. Peterson & Beech, 1967). At other times man has been seen as statistically and logically incompetent, unable to judge the simplest of probabilities or test the most straightforward of propositions (eg. Wason & Johnson-laird, 1965; Nisbett & Ross, 1980; Kahneman, Slovic & Tversky, 1982). The trend in social psychology during the 1970's was towards latter of these positions, perhaps even to a ridiculous extreme; Nisbett & Ross point out that one

of their colleagues, upon reading an early draft of their book, commented "*If we're so dumb, how come we made it to the moon*" (1980, p. 249). Similarly Edwards (1975) pointed out that unless individuals were able to assess probabilities with a certain degree of accuracy they would be unable to drive a car!

Cohen published a provocative article that again tried to turn the tide on the model of rational man. He states that "*Earlier decades, in an era of greater optimism, may well have overestimated the natural reasoning abilities of human beings. But there seems now to be a risk of underestimating them.*" (1981, p317).

Cohen goes on to show how he can categorise all experiments that claim to have demonstrated human reasoning to be invalid into one of four categories: *"cognitive illusions", "tests of intelligence or education", "misapplication of appropriate normative criteria*" and "*applications of inappropriate normative criteria*". Only the first two of these categories relate to cases where mistakes in reasoning have actually occurred, in the other two categories the mistake is on the part of the experimenters.

The categories of interest for current purposes are the first and third of these, cognitive illusions and misapplications of appropriate normative theories[1], but the others will be described first. What Cohen calls

"tests of intelligence or reasoning" refers to situations where subjects are found to be ignorant of correct rules and procedures, such as Bayes theorem or sampling theory. There is clearly no reason to expect subjects to know and understand these laws; after all they were only discovered by "great thinkers" relatively recently in the history of the human race. What is being tested for is a lack of education, leading to ignorance of the rule, or of the intelligence to apply a known rule to artificial subject matter. Cohen includes in this category those instances where an individual has insufficient knowledge of the mechanisms of the human mind and its failings -- for instance overconfidence in second-order estimates of accuracy of primary evidence. It is just as unreasonable to expect laypeople to be aware of the law of large numbers as it is to expect them to be be aware of the latest psychological findings on the mechanisms and error patterns of the mind. Cohen argues that, while failings in human reasoning that fall into this category are genuine failings that may require attention or attempts at remedy, they cannot be called irrationality, any more than ignorance of Greek mythology is irrational!

Cohen's final two categories concern errors on the part of experimenters who mistakenly claim to have discovered errors in their subjects' reasoning where none exists. Cohen gives examples of when this may occur, for instance because subjects are using the implicit rules of natural language and conversation but

experimenters are using logical statements, which have different meanings. This is an example of the experimenter misapplying normatively correct rules. Cohen's final category deals with situations where claims that experimenters are using rules in inappropriate ways. For example, he argues that there are times when prior odds should <u>not</u> be used in probability judgments. Some of his claims in this respect are controversial, and have attracted much criticism in peer commentary from the experimenters involved.

<u>"Cognitive Illusions"</u>

The category that some of Snyder & Swann's "*confirmation bias*" would best fit is that of cognitive illusions. Cohen describes these as situations in which the individual has made an error in reasoning, but an error that could be corrected with "*a few moments' prompted reflection*" (p 323). Cohen draws a metaphor between these "cognitive illusions" and visual illusions. Both of these give evidence about underlying information-processing mechanisms, but in both situations individuals can easily be made to see why they have been tricked, and can then improve their performance.

Cognitive illusions usually occur with rules that individuals are very competent at using in familiar situations, but fail to see that another, unfamiliar situation requires the same rules. For instance, the

ability to eliminate one hypothesis before moving onto another is clearly common sense in the case of "*if the soap is not in the basin it must be in the bath*", but subjects have great difficulty with it in other situations such as Mynatt, Doherty & Tweney's experiment (described earlier in this chapter).

Cohen points to another contributory factor that accounts for subjects making errors in these experiments. He accuses experimenters of deliberately contriving the situation so as to maximise the chance of errors. This is done by introducing time pressures, unfamiliar experimental materials and so on, in the same way that a conjurer relies on sl*e*ight of hand at the crucial moment to fool the audience. He concludes that, while these shortcomings in human reasoning may occur outside of the laboratory, they are probably the exception rather than the rule.

In the light of what was learned from the experiments in this thesis this description of cognitive illusions certainly seems to fit Snyder & Swann's paradigm. The list containing predominantly biased questions, the one-sided framing of the hypothesis, the audio-only communication, the judges hearing only the responses and not the questions all point to a deliberate attempt on the part of the experimenters to force the subjects into making errors in a way that would be almost impossible to occur outside of the laboratory.

Misapplication of appropriate normative criteria.

Snyder & Swann implicitly assume that the asking of unequal numbers of confirmation-seeking and disconfirmation-seeking questions is an inappropriate strategy. The preceeding section concerning Bayes' theorem earlier in this chapter, however, argued that there is no such thing as a confirmation bias in information search, only in the interpretation of information. The important thing about collecting information is that it should be diagnostic. Not only did Snyder & Swann (wrongly) assume that asking biased questions automatically leads to confirmation of the hypothesis, they also completely ignore the issues of diagnosticity and accuracy. Thus they were not only inferring that biases had occurred where this was not necessarily the case, but they were also (by providing a list of biased, and thus probably low-diagnosticity, questions) forcing sub-optimal behaviour on the subjects.

Cohen's paper generated a lot of heat from the experimenters whom he was criticising, but not much consensus. Perhaps the reason for this was best summed up in one of those peer commentaries, by Evans & Pollard (1981). They point out that Cohen's central argument, that neither the existing literature nor any possible experiment into deductive or probablistic reasoning can ever establish faulty competence, is an irrefutable philosophical stance "*of little practical relevance to the scientific study of human reasoning*" (p. 335). If rationality is defined in such a way that no adult human being can ever be irrational (as Cohen clearly does), then the definition of rationality is most unhelpful.

Baron's book entitled "*Rationality and Intelligence*" (1985) manages to avoid these philosophical problems by defining rationality differently, in a way that is more useful for evaluating human cognitive performance. In Chapter 1 of his book Baron explains that rationality is a property of thinking that may be present or absent in varying proportion. Rational thinking is defined as the following of a good model of decision making or belief formation. The model is good if it maximises the chance of conforming to the normatively correct model. It is not necessarily the same as the normatively correct model, because it takes into account constraints such as the amount of time and effort to be invested in thinking. A rational process for thinking is thus the one that is most likely to lead to the desired outcome within the constraints on the

decision-making process. While it is still possible to arrive at the optimal decision with an irrational strategy (for instance, through luck) and possible to arrive at a bad outcome after a rational decision (for instance through unforseeable circumstances), the rational decision process is the one that is most likely to succeed.

Baron also argues that it is inadequate to consider rationality in isolation, but that it needs to be considered in the context of intelligence, effectiveness, goals and so on. Thus Baron does not limit his definition of rationality to systematic biases, but also to non-systematic errors. Furthermore, it can be inferred from Baron's position that if two decision-making strategies were equally likely to lead to the correct answer, but one method took more time and effort, then that strategy would be less rational than the other. Perfect rationality is thus almost never achieved, but serves as a criterion against which human performance can be measured.

One other notable approach to the question of errors and biases is Kruglanski's theory of lay epistomology (Kruglanski & Ajzen, 1983; Kruglanski, Baldwin & Towson, 1983). In its implications, Kruglanski's theory is entirely compatible with Baron's, that biases and errors are a necessary and unavoidable consequence of the shortcuts we need to take in our thinking process.

Kruglanski sees lay epistomology to be the unifying principle involved in many disparate areas of social psychology from cognitive consistency through attribution theory to judgment under uncertainty. The lay epistemic process consists of testing hypotheses with a series of "*only if x then y*" deductions. New hypotheses are generated when a deduction does not support the old hypothesis, and so the process continues.

Kruglanski also points out that any finite set of observations is consistent with an infinite set of hypotheses. Therefore we can never fully finish the epistemic process when the belief is held with absolute certainty. At some point it has to be decided that the process will be "frozen". The process may then be "unfrozen" at a later point in time for any one of a number of reasons, for instance a stronger motivation to be accurate or because new information becomes available.

While both Kruglanski and Baron share the view that biases and errors are inevitable, they also both believe that people's thinking can be made more accurate by educational intervention. For instance, by alerting people to the availability heuristic or regression to the mean, people will be better able to avoid falling prey to some of the worst errors.

A different and novel approach to the rationality issue is to look at all the sorts of tasks that people are required to make and divide them into those that people are good at and those where they are outperformed by computers (Topmiller, 1979). Interestingly, in an analysis of this sort, computers would probably be judged far superior to subjects according to the rules of Snyder & Swann's paradigm or the Bayesian and diagnosticity experiments, but the evidence in this thesis points to the fact that in actual social interactions present-day computer programmes would be totally outperformed by people.

Biases and Errors

It is worth, at this point, defining exactly what is meant by the terms error and bias, as they are often used interchangably in the literature. An error is a deviation between an actual value and the predicted or estimated value. Its direction is not predictable. A bias also implies an error, but one that is "*system-atic, consistent and predictable*" (Schneider, Hastorf & Ellsworth, p.226, 1979). For instance, when assessing the personality of others, if I was often completely wrong (for example calling extraverts introverted and neurotics stable) then my judgments would include much error. If, though, there was some pattern to these errors then they could be called biases. If, for instance, I rated everyone as more extraverted than they actually were then I would have an

"*extravert bias*". Similarly if I rated everyone as conforming to the hypothesis I was testing, then I would have a "*confirmation bias*". While this is the most common use of the term bias in social cognition, see Schneider, Hastorf & Ellsworth, (1979, pp 225-226) for other uses of the term.

Hypothesis testing and errors and biases

Which of the behaviours and judgments observed in the hypothesis testing experiments presented here and elsewhere are evidence of irrational information search and processing on the part of subjects?

Snyder makes his position on errors and biases explicit in two separate articles (Snyder, 1981; Snyder & Gangestad, 1981). He argues that nothing observed in his experiments can be called an error or a bias. His argument goes as follows. As it happened, subjects chose to look for confirmatory evidence, and found it, which led them to conclude that their hypotheses were valid. If, though, they had chosen to search for disconfirmatory evidence, they would no doubt have found it, and concluded that their initial hypotheses were invalid. The other option open to subjects would have been to use an "equal opportunities" strategy, searching for confirmatory and disconfirmatory evidence in equal proportion. And if they had done this, they would have concluded that anyone they interviewed was half introvert and half extraverted. Since, of these three

possible strategies, not one of them is any better than any other, it would be wrong to call any or all of them irrational or biased. This, says Snyder, shows the difference between the social and the physical world. Any interactive hypothesis testing strategy in the social world is <u>reactive</u>, and thus no strategy is better than any other.

Snyder & Gangestad take this argument further and say that even in the "historical" hypothesis testing experiment (Snyder & Cantor, 1979) it would be incorrect to call the automatic confirming of hypotheses an error. After all, they say, we know that people are changed by the expectations and beliefs others have of them (Snyder, Tanke & Berscheid, 1977). So, even if an introverted person was taken on for a job which required an outgoing sort of person, that person might soon become outgoing, at least in their working lives.

It will be argued here that this argument is flawed and untenable. It is tantamount to saying that there is <u>no</u> social reality, that there are <u>no</u> enduring differences between people except in the way that others see them. Taken to its logical extreme, Snyder's argument would suggest that job interviews are completely redundant as a way of selecting people suitable for a job, as are clinical psychologists who attempt to diagnose personality disorders. Even personality questionnaires like the *EPI* or the *MMPI* are totally useless, because if the questions had been phrased in a

different way, then everyone could have been an introvert, an extravert or half way between! Surely not even the most radical situationalist would go that far!

The first two experiments presented in this thesis were able to show that the interviewers, even asking Snyder's "biased" questions could determine quite accurately the extraversion of the targets, as measured by the EPI.

Bias in Information Search

Firstly, consider the information search stage of the hypothesis testing process. To what extent can the asking of only Snyder's confirmation-seeking questions be said to be biased or sub-optimal as a strategy? It has already been argued that there is no such thing as biased information search (Fischhoff & Beyth-Marom, 1983), only biased interpretation of that evidence. So, it is perhaps misleading to call the type of questions used by Snyder biased, but it is meant to demonstrate that they tend to elicit an unrepresentative (non-random) set of information. There are three things that subjects could do having asked a series of, say, biased extravert questions:

1/ They could realise that the information they had collected was unrepresentative, and thus ignore it completely.

2/ They could realise that the information thay had collected was unrepresentative and attempt to take this into account when interpreting that information, working out to what extent the information collected is evidence of the target's extraverted disposition and to what extent it is representative of the questions asked.

or

3/ They could ignore the fact that the information that they had collected was unrepresentative and treat it as if it were a fair representation of the target's disposition.

What of errors and biases in these three situations? In the first situation, the subject has wasted his or her time, because the information collected was not used in the judgment. Because of this, his judgments about the target are likely to be inaccurate, since there will be little or no remaining evidence to base them upon. Therefore, the judgment will be high in error, but not biased, in general. The decision-making process would have been inefficient, but in the case of experiments which used only Snyder's type of biased questions, the subject would probably have been justified in feeling that the inefficiency had been forced upon them by the experimental materials.

In the second situation, any bias in the assessment of the target's personality would be at the stage of the interpretation of the information, not at the stage of collection. However, the task of interpreting the information may have been made needlessly complex; if a representative set of data had been collected the interpretation would have been more straightforward. This being the case, while the method of information search could not be said to be the sole cause of biases and errors, it might be a contributory factor. Similarly, if the question that they had asked had been low in diagnosticity because they sought out only confirmatory evidence, then this could also have led to inaccuracies (Klayman & Ha, 1986).

In the third possibility the interviewers' perceptions of the target will clearly be in error, having failed to take into consideration a relevant factor in the interpretation of the data, the unrepresentativeness of that data.

From this analysis it would seem that any inaccuracies that might affect subjects' judgments, would at least in part be caused by the rather odd set of questions used in the experiments, but this is not inevitable. In order for there to be a confirmation bias, it needs to occur at the information processing stage, not the information search stage. While this discussion is somewhat irrelevant given that subjects do not ask biased questions anyway, it serves to make the more general point that even when inferences are made from an unrepresentative set of data, those inferences need not be biased. It is possible for the unrepresentativeness of the information to be recognised and taken into account. While there is plenty of evidence to suggest that the representativeness of samples is often ignored in human reasoning (particularly when the availability heuristic is used) it is not always the case.

This discussion, as well as accepting Snyder's assumptions about the types of questions asked, also accepts Snyder's model of the way in which people collect and interpret information. Trope & Bassok (1982, 1983) also assume that the primary information of

interest to the interviewer is the content of the target's response. Yet the debriefing of subjects in the experiments reported in chapters 6 and 7 clearly showed that the hypothesis testers considered that style was at least as important. When asked why they wanted to ask questions such as "*What is your opinion of marriage?*" they readily justified them with statements like "*It will get her talking*" or "*I'd start with that one to open him up even if he is very shy*". This concentration on the non-verbal or more subtle aspects of conversation are even more likely in face to face interactions than in paper and pencil or audio-only situations (Short, Williams & Christie, 1976). Thus, even questions that generate unrepresentative answers with respect to content may still generate a style more representative of the target's normal behaviour. While the subjects' claiming that these things were important to them does not necessarily mean that they are the important variables in the subjects decision making process (Nisbett & Wilson, 1977) it is certainly a point that should considered further before going on to build increasingly elaborate models of hypothesis testing processes.

Random Error

Whereas Snyder claims that the concept of bias is meaningless in testing hypotheses using social knowledge, he does not consider the possibility and implications of error. It is possible that, even if

there were no systematic confirmatory deviations in hypothesis testers' judgments, they might still be wildly inaccurate due to random error. This is an important omission in any full understanding of social hypothesis testing. The experiments in chapters 3 and 4 of this thesis go some of the way towards answering this question by comparing the interviewers' perceptions of the targets with the actual targets' dispositions measured using standard psychological tests (and, incidentally, finding a high correspondence). Trope & Bassok's experiments also address the question of error. The issue of diagnosticity is inextricably linked with error in this instance. The greater the diagnosticity of the questions asked, the smaller the error is likely to be in the final judgment (or, put another way, the nearer the subjective posterior odds are likely to be to the objective truth).

"Practical Rationality"

There is another way in which rational considerations could affect the choice of questions. The rationality of a decision or action can only be judged in the light of the goals of the individual. So far it has been assumed that the interviewers in these experiments had only one goal, to determine the validity of the hypothesis. This is undoubtably an over-simplification of the interviewers' role. They were also probably trying to simultaneously be polite to the target, please the experimenter, satisfy their own

curiosity, manage the impression that the target formed of them and so on.

These considerations are what Mortimore calls "*practical rationality*" as opposed to "*epistemic rationality*" (Abelson, 1976), and can account for some of the other phenomena noticed in question selection such as the avoidance of the introvert questions, especially when testing for extraversion. This points to another way in which experimenters may mis-interpret the behaviour of subjects, by failing to identify their goals properly. While it is important not to take this approach too far and try to interpret all behaviour as rational by re-specifying the subjects' goals with the benefit of hindsight (Fischhoff, 1982), it would be over-simplistic to treat subjects in conversation as being interested only in finding out about the other person.

Suggestions for further research on Hypothesis Testing in Social Interaction

Despite the failure to find any substantial evidence of a predisposition to use interactions to confirm hypotheses on the part of subjects, it would be wrong to conclude that these effects cannot occur. As Fischhoff (1982) points out, it can be just as misleading to paint too rosy a picture of human rationality. What is clear, however, is that if the self-confirming hypothesis phenomenon is to occur in social interaction, it is not going to be in the way suggested by Snyder & Swann (1978).

There is evidence from other paradigms that the type of questions are asked can have a marked effect on the answers received, and perhaps these could be pursued to see if they can be applied to social interaction. For instance, there is good evidence that asking different sorts of questions to elicit eyewitness testomonies can have marked effects on recall of facts and impressions (Loftus, 1975; Lipton, 1977; Marshall, Marquis & Oscamp, 1971). Another related phenomenon of interest is the finding that subtle changes in the wording of questions in attitude inventories (Rois er, 1974) or public opinion polls (Orenstein & Phillips, 1978)can also have a marked effect on the responses to those questions.

The arguments presented so far in the thesis carry several implications about the way in which research into hypothesis testing in social interaction should proceed. These suggestions will probably apply equally to other research that may be conducted into hypothesis testing processes in social interaction using methodologies very different to Snyder & Swann's. These suggestions will be discussed in turn.

Artificiality

It is clear from the arguments presented in this thesis that several of the techniques used in Snyder & Swann's "paradigm setting" experiments were so artificial (and needlessly so) as to give them no external validity or generalisability. Examples of this were the choice of questions, the audio-only communication, the use of naive judges instead of the interviewers themselves, the selection of questions during rather than before the experiment, and so on. It is not being argued here that virtually all laboratory experiments are so artificial as to be incapable of leading to an understanding of human cognition and human behaviour (as Harré & Secord (1972) do at times). There is nothing in the present thesis to suggest that laboratory experiments *per se* cannot be of use in the scientific study of Social Psychology. Rather, it is argued here that when the procedures used in experiments are so removed from those normally employed in natural environments, their generalisability will be very low. There is no

need to use artificial procedures just because a high degree of control is required; The experiment in chapter 8 of this thesis and the experiments by Forgas into person perception in real groups (1976, 1978, 1981(a&b), 1982, 1983) are good examples of this, and there are plenty of others. All that is required is some additional, and perhaps creative, thinking by the experimenter, and maybe also some observation of real people in the situation that one wants to generalise to. Reading Snyder & Swann's experiment one could be forgiven for thinking that neither of them had actually seen two people talking to each other, let alone a job interview or any other situation where hypotheses might be being tested!

The framing of a task

Earlier in this chapter it has already been argued that Snyder's way of framing hypotheses for subjects may be completely different from the way in which subjects set themselves hypotheses to test. In order to investigate the errors and biases that subjects are prone to in testing their hypotheses, it is essential to study a similar process to that which one is generalising to. In this case it will require some further work in order to investigate the way in which hypotheses are naturally framed.

This is not as simple as it may sound, and no simple observation or experiment is being recommended here. Perhaps it will require listening to natural

speech as people ask each other questions. Perhaps it will require asking subjects questions like "*suppose you were to tell someone how to interview people for the job of salesperson. What would you tell them to do?*". It is difficult problem for future research to tap, but it will be necessary if lay hypothesis testing is to be investigated further. It will not suffice to continue with the current assumptions about how people test hypotheses; not are they only assumptions but, it is argued here, implausible assumptions to boot.

Replication

It is often said that to be sure of an effect in the social sciences one must make sure that it can be replicated, preferably by a different experimenter and with slightly different experimental materials. There have been too many effects reported that have failed to replicate (See, for instance, Morley, 1978 or Amir & Sharon, 1984). Not only might an effect be a *type one* error, but it also may have been an artefact of the particular subjects used, the particular characteritics of the experimenter, numerical errors, fraud or any one of a multitude of other pitfalls of experiments (See Barber, 1976 for a full discussion). Replication is particularly true of experiments that try to generalise from one sample to another, in this case from under-graduates to job interviewers and other professionals.

Yet, a large and still growing part of the literature in social cognition has rested on a few very

weak links in the chain of events in hypothesis testing such as Swann's PhD thesis and the judging of the targets' responses in Snyder & Swann's second investigation (1978). Other links in the chain have been replicated over and over many times, such as the question selection task. Given that a chain is only as strong as its weakest link, this is a very inefficient way to do research. In this case (as this thesis has demonstrated) those weak links have caused the whole chain to break!

The whole is not the sum of the parts

Snyder's approach to hypothesis testing, and also to self-fulfilling prophecies, has been that they are causal chains made up of many links that can be studied seperately. While this is a good way to get a good understanding of each of those stages, demonstrating that a bias can occur at a particular stage is not the same as demonstrating that the same bias will occur when embedded in the whole process. There have been several demonstrations to show that the shortcomings in judgment may be observed in one situation but not in another similar situation. There is nothing all-pervasive about, for instance, the availability heuristic or the under-utilisation of base rates. As Kruglanski and others have shown (Kruglanski & Ajzen, 1983; Ajzen, 1977) the biases can easily be manipulated by small changes in the situation. So, a general knowledge that, for instance, individuals tend to underestimate the

influence of situational forces in determining the behaviour of others does not mean that individuals will fall prey to this bias in all situations. Snyder continually makes inferences of this sort, inferring that because judges perceived confirmation in the target's behaviour, the interviewers would also interpret that behaviour as evidence confirming the hypothesis, or inferring that, because interviewers tend to conclude that their hypotheses are confirmed in one situation (Swann, 1978) they will also confirm their hypotheses in another, completely different situation (Snyder & Swann, 1978).

Dynamic or static process?

Snyder's treatment of hypothesis testing is as if it were a completely static process. The hypothesis tester receives the hypothesis, selects questions to ask, asks those questions and then interprets the responses. Hypothesis testing is thus characterised as a uni-directional linear progression of stages, each one leading to the next.

Person perception is rarely like this. Individuals may set hypotheses, then get some evidence, which prompts them to either change the hypothesis or search for a different set of information. The first experiment (Chapter 4) demonstrated this nicely, that people modify their search for information as they proceed through the task. To use a metaphor, if Snyder

has represented hypothesis testing as being like the firing of a bullet in a specific direction, real life person perception can be seen as more like the flight of a guided missile. By making continual changes as the target is approached, much greater accuracy is possible that a "one off" aim and fire.

Hogarth (1981) has made a similar point. Because most research has focused on discrete incidents, he argues that the role of feedback in decision making has been virtually ignored.

As well as designing experiments to allow for feedback and a modifying of strategy, other changes would be desirable to better understand the processes involved. If the interviewer's perception of the target is to be seen as continually in flux, but presumably "homing in" on a more stable perception of the target, then it would be very informative to follow the progress of that representation over the duration of the interaction (or, even better, interaction<u>s</u>).

The most obvious way to do this would be to stop the interaction at, say, two minute intervals and each time measure the interviewer's impression of the target. This, however, may have the problem of interfering with the natural course of the interaction. Experimenters may then have to settle for a slightly less powerful and revealing design. Many interactions could be set up using the same expectations, hypotheses and conditions.

Some would be stopped after two minutes, some after four, some after six, and so on. The changes in accuracy and bias could still be measured at several points in time, but it would not be truly longitudinal as the first one would.

Decisions taken socially, not individually

After reading the research presented here and elsewhere on the testing of hypotheses about other people it is perhaps appropriate to conclude that it is too early to start making inferences from it to the real world. There is, though, one important conclusion that would merit further research.

The literatures on self-fulfilling prophecies and self-confirming hypotheses both point to the power of expectations and hypotheses to perpetuate erroneous stereotypes and prejudices about other people. The experiments and analyses presented here suggest that these effects might be very unlikely to occur when there is social interaction between perceivers and targets, but much more likely when there is no direct contact during the decision-making (as in the experiment in Chapter 8). The conclusion that interaction between stereotyped groups will often be counter-productive may have been over emphasised in Social Psychology; social processes may be the factor that prevent individuals from falling prey to the limitations of their individual information-processing systems. If this is the case

then Social Psychology will need to develop methods of studying judgment under uncertainty as a social not an individual phenomenon.

Belief Perseverance

One finding in social cognition is that individuals will continue to hold a belief, even when all of the evidence that lead to that belief is completely discredited in as vivid a manner as possible (Ross, Lepper & Hubbard, 1975). Snyder & Swann's work on hypothesis testing in social interaction may have developed an equivalent inertia in the social cognition literature itself, perhaps because the initial findings were so plausible, particularly if one doesn't look too closely at the methods they used. This is, perhaps, the reason why Snyder's work on hypothesis testing continues to be an accepted part of the wisdom of Social Psychology even after all of the contrary evidence that has been published, a fraction of which is enough to discredit it completely. Take, for example, Fiske & Taylor who, after reviewing some of the literature critical of Snyder & Swann, state (referring to social interaction) that "*There seems little doubt that confirmatory hypothesis testing occurs, but how robust it is remains an issue.*" (1984, p. 387). Even in 1986 Snyder & Swann's initial experiment is still cited uncritically in prestigeous books and journals such as the *Annual Review of Psychology* (Miller & Turnbull).

I have rarely seen a better demonstration of belief perseverance, but have yet to be shown a convincing demonstration of self-confirming hypotheses <u>in social interaction!</u> And no doubt even when social psychologists themselves have managed to resolve this issue, the ghost of the self-confirming hypothesis will continue to haunt applied psychology for a good many more years.

Summary and Conclusions

1/ This chapter started by considering several experiments on hypothesis testing that used a Bayesian conceptual framework (eg Trope & Bassok, 1982, 1983). It was concluded that the findings of the experiments in chapters four to seven (*ibid*) complemented their overall conclusions that, when it comes to social interaction, individuals are fairly good intuitive Bayesians. However, like Snyder & Swann's experiments, those experiments also used a very simplistic model of social interaction which the experiments presented in this thesis demonstrated is probably inadequate when investigating hypothesis testing processes. Nevertheless, Bayes' Theorem provides a very clear conceptual framework for analysing hypothesis testing and if, for instance, Snyder & Swann had used it in constructing their experiments they could have avoided some of their conceptual errors and been better able to understand the contribution of both information search and information processing to confirmation biases.

2/ The Hypothesis-testing paradigm was then compared and contrasted to the attribution theory paradigm. It was argued that these two paradigms shared a lot in common, but for some reason they had tended to co-exist rather than support each other. It was then argued that a lot could be gained by considering the similarities and differences between these two fields. Several criticisms of attribution theory were shown to

be very relevant to the hypothesis testing paradigm. For instance, in order to understand hypothesis testing processes in everyday life it will be necessary to understand the way in which individuals frame hypothesis testing tasks. In addition, the hypothesis testing paradigm (Like the attribution paradigm) assumes a high level of information processing in everyday inter-actions; this may be misleading.

3/ The behaviour of the "interviewers" in hypothesis testing experiments was evaluated in the light of the literature on human rationality. It was concluded that some of the decisions such as the choice of questions that had been labeled by some researchers as biased could, in fact, have been entirely consistent with a normatively correct strategy. Where real biases may have occurred they were probably forced upon the subjects by the artificiality of the experimental materials rather than being symptomatic of some underlying flaw in human reasoning.

4/ Snyder's position (1981; Snyder & Gangestad, 1981) that it is not possible to call a decision based on social as opposed to physical reality biased or inaccurate was argued to be wrong. It is tantamount to saying that there are no real differences between people.

5/ It was concluded that, while attempts to find experimental evidence of a confirmation bias in testing

hypotheses using social interaction had not been successful in this thesis, this was not to say that they could not exist. Suggestions were made for future research into hypothesis testing and other studies of the social interaction processes. It was argued that a better understanding of the social interaction process is needed before imposing constraints on experimental subjects, that the research should be more "wholistic" rather than dividing processes up into all of their constituent parts, that important effects should be tested for replication and that social interaction should be considered as a dynamic process with feedback rather than a static one-off decision. It was also suggested that decisions taken in social interactions may be less prone to error than judgments made individually. The individualistic emphasis of the *judgment under uncertainty* literature may contribute to the bleak picture it paints of human rationality.

6/ Finally concern was expressed that, while the evidence that led to our knowledge of a confirmation bias in hypothesis testing has been completely discredited, the effect is still accepted uncritically in the psychology literature. Most of the empirical and conceptual evidence in this thesis points to the fact that social interactions (as opposed to individual cognitive processes) cause erroneous hypotheses and expectations to be rejected rather than perpetuated.

References

A

Abelson, R.P. Social Psychology's rational man. In S.I. Benn & G.W.Mortimore (Eds.) Rationality and the Social Sciences. London: Routledge and Kegan Paul, 1976.

Abelson, R.P. Script processing in attitude formation and decision making. In J.S. Carroll & J.W. Payne (Eds.) Cognition and Social Behavior. Hillsdale, New Jersey: Earlbaum, 1976(b).

Andersen, S.M. & Bem, S.L. Sex Typing and androgeny in dyadic interaction: Individual differences in responsiveness to physical attractiveness. Journal of Personality and Social Psychology, 1982, 43, 22-34.

Amir, Y. & Sharon, I. Are Social-Psychological Laws Cross-Culturally Valid? Unpublished report, Bar-Ilan University, Israel, 1984. (Cited in Jahoda, 1986).

Antaki, C. Ordinary Explanation, Attribution Theory and Verbatim Accounts. Unpublished doctoral dissertation, University of Sheffield, 1982. (Cited in Eiser, 1983.)

Argyle, M. & McHenry, R. Do Spectacles really affect Judgements of Intelligence? British Journal of Social and Clinical Psychology, 1971, 10, 27-29.

Asch, S. Forming impressions of personality. Journal of Abnormal and Social Psychology, 1946, 41, 258-290.

B

Barber, T.X. Pitfalls in human research: Ten pivotal points. New York: Pergamon Press, 1976.

Baron, J. Rationality and Intelligence. Cambridge: Cambridge University Press, 1985.

Berman, J.S., Read, S.J. & Kenny, D.A. Processing inconsistent Social Information. Journal of Personality and Social Psychology, 1983, 45, 1211-1224.

Berscheid, E. & Walster, E. Physical attractiveness. In L. Berkowitz (Ed.) Advances in Experimental Social Psychology (Vol. 7). New York: Academic Press, 1974.

Beyth-Marom, R. & Fischhoff, B. Diagnosticity and Pseudodiagnosticity. Journal of Personality and Social Psychology, in press.

References. page 342

Burchell, B.J. Self-Fulfilling Prophecies verses Self-Confirming Hypotheses: The differences are larger than the similarities. Paper presented to the British Psychological Society (Social Psychological Section) Annual Conference, Oxford, 1984.

Burchell, B.J. Information search and the testing of hypotheses about other people. Questioning Exchange, 1987, in press.

Burchell, B. J. and Morley, I.E. Is there a confirmatory Bias? A critique of Snyder & Swann (1978). Paper presented to the British Psychological Society (Social Psychological Section) Annual Conference, Sheffield, 1983.

C

Carver, C.S. & de la Garza, N.H. Schema-Guided Information Search in the Stereotyping of the Elderly. Paper presented to the Eastern Psychological Association, Baltimore, 1982.

Chapman, A.J. & Jones, D.M. (Eds.) Models of Man. Leicester: British Psychological Society, 1980.

Chapman, L.J. & Chapman, J.P. Genesis of popular but erroneous diagnostic beliefs. Journal of Abnormal Psychology, 1967, 72, 193-204.

Chapman, L.J. & Chapman, J.P. Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. Journal of Abnormal Psychology, 1967, 74, 271-280.

Clark, H.H. & Clark, E.V. Psychology and Language: An introduction to psycholinguistics. New York: Harcourt Brace Jovanovich, 1977.

Cohen, L.J. Can human irrationality be experimentally demonstrated? The Behavioural and Brain Sciences, 1981, 4, 317-370.

Cooper, W.H. Controlling a confirmatory bias in interpersonal hypothesis testing. Unpublished Manuscript, Queen's University, Ontario, 1982.

Crano, W.D. & Mellon, P.M. Causal influence of teachers' expectations on children's academic performance: A cross-lagged panel analysis. Journal of Educational Psychology, 1978, 70, 39-49.

Crocker, J. Biased questions in Judgment of Covariance studies. Personality and Social Psychology Bulletin. 1982, 8, 214-220.

References. page 343

Darley, J.M. & Fazio, R.H. Expectancy Confirmation Processes Arising in the Social Interaction Sequence. American Psychologist, 1980, 35, 867-881.

Darley, J.M. & Gross, P.H. A Hypothesis-confirming Bias in Labeling Effects. Journal of Personality and Social Psychology, 1983, 44, 20-33.

Dion, K.K. & Berscheid, E. Physical attractiveness and peer perception among children. Sociometry, 1974, 37, 1-12.

Dion, K.K., Berscheid, E. & Walster, E. What is beautiful is good. Journal of Personality and Social Psychology, 1972, 24, 285-290.

Doherty, M.E., Mynatt, C.R., Tweney, R.D. & Schiavo, M.D. Pseudodiagnosticity. Acta Psychologica, 1979, 43, 111-121.

E

Edwards, W. Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. Journal of Mathematical Psychology, 1965, 2, 312-329.

Edwards, W. Comment. Journal of the American Statistical Association, 1975, 70, 291-293.

Edwards, W., Lindman, H. & Savage, L.J. Bayesian Statistical Inference for Psychological Research. Psychological Review, 1963, 70, 193-242.

Eiser, J.R. Cognitive Social Psychology. London: Mcgraw-Hill, 1980.

Eiser, J.R. From Attributions to Behaviour. In M. Hewstone (Ed), Attribution Theory: Social and Functional Extensions. Oxford: Basil Blackwell, 1983.

Erdelyi, M.H. A new look at the New Look. Perceptual defence and vigilance. Psychological Review, 1974, 81, 1-25.

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

Evans, J.St. B.T. Thinking: Experiential and information processing approaches. In G. Claxton (Ed.), Cognitive Psychology: New Directions. London: Routledge & Kegan Paul, 1980.

Evans, J.St. B.T. & Pollard, P. On defining rationality unreasonably. The Behavioural and Brain Sciences, 1981, 4, 335-336. (in commentary on Cohen, 1981)

References. page 344

Fischhoff, B. Attribution theory and Judgement under Uncertainty. In J.H. Harvey, W.Ickes & R.F. Kidd (Eds.) New directions in Attribution Research (Vol. 1). Hillsdale, New Jersey: Earlbaum,1976.

Fischhoff, B. Lattitudes and Platditudes: How much credit do peo£le deserve? in G. Ungson & D.Braunstein (Eds.) Decision Making: An Interdisciplinary Enquiry. New York: Kent, 1982.

Fischhoff, B. & Beyth-Meyrom, R. Hypothesis-Evaluation from a Bayesian Perspective. Psychological Review, 1983, 90, 239-260.

Fiske, S.T. & Taylor, S.E. Social Cognition, London: Addison-Wesley, 1984.

Forgas, J.P. The perceptions of social episodes: Categorical and dimensional representations in two social milieus. Journal of Personality and Social Psychology, 1976, 33, 199-209.

Forgas, J.P. Social episodes and social structure in an academic setting: The social environment of an intact group. Journal of Experimental Social Psychology, 1978, 14, 434-448.

Forgas, J.P. Social Cognition: Perspectives on everyday understanding. New York: Academic Press, 1981(a).

Forgas, J.P. What is social about social cognition? In J.P. Forgas (Ed.) Social Cognition: Perspectives on everyday understanding. New York: Academic Press, 1981(b).

Forgas, J.P. Affective and emotional influences on episode representations. In J.P. Forgas (Ed.) Social Cognition: Perspectives on everyday understanding. New York: Academic Press, 1981(c).

Forgas, J.P. Epilogue: Everyday understanding and social cognition. In J.P. Forgas (Ed.) Social Cognition: Perspectives on everyday understanding. New York: Academic Press, 1981(d).

Forgas, J.P. Episode cognition: Internal representations of interaction routines. In L. Berkowitz (Ed.) Advances in Experimental Social Psychology (Vol. 15). New York: Academic Press, 1982.

Forgas, J.P. What is social about social cognition? British Journal of Social Psychology, 1983, 22, 129-144.

Forgas, J.P. Person prototypes and cultural salience: The role of cognitive and cultural factors in impression formation. British Journal of Social Psychology, 1985, 24, 3-17.

References. page 345

Fransella, F. Man as scientist. In A.J. Chapman & D.M. Jones (Eds.), Models of Man. Leicester: British Psychological Society, 1980.

G

Goldman, W. & Lewis, P. Beautiful is good: Evidence that the physically attbactive are more socially skillful. Journal of Experimental Social Psychology, 1977, 13, 125-130.

Gollob, H.F., Rossman, B.B. & Abelson, R.P.. Social inference as a function of the number of instances and consistency of information presented. Journal of Personality and Social Psychology, 1973, 27, 19-33.

Gorman, C.D., Clover, W.H. & Doherty, M.E. Can we learn anything about interviewing real people from "interviews" of paper people? Two studies of the external validity of a paradigm. Organisational Behavior and Human Performance. 1978, 22, 165-192.

Greenwald, A.G. The totalitarian ego: Fabrication and revision of personal history. American Psychologist, 1980, 35 603-618.

H

Hamilton, D.L. A cognitive-attributional analysis of stereotyping. In L. Berkowitz (Ed.) Advances in Experimental Social Psychology (Vol. 12). New York: Academic Press, 1979.

Hamilton, D.L. Illusory correlation as a basis for stereotyping. In D.L.Hamilton (Ed.), Cognitive Processes in Stereotyping and Ingroup Behavior. New Jersey: Earlbaum, 1981(a).

Hamilton, D.L. Cognitive Processes in Stereotyping and Ingroup Behavior. New Jersey: Earlbaum, 1981(b).

Hamilton, D.L. Stereotyping and intergroup behavior: Some thoughts on the cognitive approach. In D.L.Hamilton (Ed.), Cognitive Processes in Stereotyping and Ingroup Behavior. New Jersey: Earlbaum, 1981(c).

Hamilton, D.L. & Gifford, R.K. Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. Journal of Experimental Social Psychology, 1976, 12, 392-407.

Hamilton, D.L. & Rose, T. Illusory correlations and maintainance of stereotypic beliefs. Journal of Personality and Social Psychology, 1980, 39, 832-845.

Hansun, R.D. Commonsense attribution. Journal of Personality and Social Psychology, 1980, 39, 996-1009.

References. page 346

Harré, R. Rituals, retoric and social cognition. In J.P. Forgas (Ed.) Social Cognition: Perspectives on everyday understanding. New York: Academic Press, 1981.

Harré, R. & Secord, P.F. The Explanation of Social Behaviour. Oxford: Blackwell, 1972.

Harris, B. & Harvey, J. Attribution theory: From phenomenal causality to the intuitive social scientist and beyond. In C. Antaki (Ed.) The Psychology of Ordinary Explanations of Social Behaviour. London: Academic Press, 1981.

Hastie, R. & Kumar, P.A. Person memory: Personality traits as organising principles in memory for behaviours. Journal of Personality and Social Psychology, 1979, 37, 25-38.

Heider, F. The Psychology of Interpersonal Relations. New York, Wiley, 1958.

Hilton, J.L. & Darley, J.M. Constructing other persons: A limit on the effect. Journal of Experimental Social Psychology, 1985, 21, 1-18.

Hogarth, R.M. Beyond discrete biases. Functional and dysfunctional aspects of judgmental heuristics. Psychological Bulletin, 1981, 90, 197-217.

Hov land, C.I. & Weiss, W. Transmission of information concerning concepts through positive and negative instances. Journal of Experimental Psychology, 1953, 45, 175-182.

I

Isen, A.M. & Levin, P.F. The effect of feeling good on helping: Cookies and kindness. Journal of Personality and Social Psychology, 1972, 21, 384-388.

J

Jahoda, G. Nature, culture and social psychology. European Journal of Social Psychology, 1986, 16, 17-30.

Jenkins, H.M. & Ward, W.C. Judgment of contingency between responses and outcomes. Psychological Monographs, 1965, 79.

Jennings, D., Amabile, T.M. & Ross, L. Informal covariance Assessment: Data based verses theory based judgments. In D. Kahneman, P. Slovic, & A. Tversky, (Eds.) Judgment Under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press, 1982.

References. page 347

rvis, R. Perception and Misperception in International Politics. New Jersey: Princetown University Press, 1976.

Johnson, E.J. & Tversky, A. Affect generalisation and the perception of risk. Journal of Personality and Social Psychology, 1983, 45, 20-31.

Johnson, T.J., Feigenbaum, R. & Weiby, M. Some determinants and consequences of teachers' perception of causality. Journal of Educational Psychology, 1964, 55, 237-246.

Johnson-Laird, P.N., Legrenzi, P. & Sonino Legrenzi, M. Reasoning and a sense of reality. British Journal of Psychology, 1972, 65, 395-400.

Jones, E.E. & Davis, K.E. From acts to dispositions: The attribution process in social perception. In L. Berkowitz (Ed), Advances in Experimental Social Psychology (Vol 2). New York, Academic Press, 1965.

Jones, E.E. & Nisbett, R.E. The actor and the observer: divergent perceptions of the causes of behavior. In E.E. Jones, D.E. Kanouse, H.H. Kelley, R.E. Nisbett, S. Valins and B. Weiner, (Eds.), Attribution: Perceiving the Causes of Behavior. New York: General Learning Press, 1971.

Jones, R.E. Self-Fulfilling Prophecies: Social, Psychological and Physiological Effects of Expectancies. Hillsdale, New Jersey,: Lawrence Earlbaum Associates, 1977.

K

Kahneman, D. & Tversky, A. Subjective probability: A judgment of representativeness. Cognitive Psychology, 1972, 3, 430-454.

Kahneman, D. & Tversky, A. On the psychology of prediction. Psychological Review, 1973, 80, 237-251.

Kahneman, D. & Tversky, A. The psychology of preferences. Scientific American, 1982, 136-142.

Kahneman, D., Slovic, P. & Tversky, A. Judgment Under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press, 1982.

Kelley, H.H. Attribution theory in social psychology. In D. Levine (Ed), Nebraska Symposium on Motivation, Lincoln, Nebraska: University of Nebraska Press, 1967.

Kelley, H.H. Causal Schema and the attribution process. In E.E. Jones et al (Eds), Attribution: Perceiving the causes of behaviour. Morristown, New Jersey: General Learning Press, 19722.

References. page 348

...lley, H.H. & Stahelski, A.J. Social interaction basis of cooperators' and compeditors' beliefs about others. Journal of Personality and Social Psychology, 1970, 16, 66-91.

Kelly, G. The Psychology of personal Constructs (2 Vols.). New York: Norton, 1955.

Keisler, S. & Sproull, L. Mangerial responses to changing environments: Perspectives on problem sensing from social cognition. Administrative Science Quarterly, 1982, 27, 548-570.

Keppel, G. Design and Analysis: A Researcher's Handbook. New Jersey: Prentice-Hall, 1973.

Kindler, D.R. & Weiss, J.A. In lieu of rationality: Psychological perspectives on foreign policy decision making. Journal of Conflict Resolution, 1978, 22, 707-735.

Klayman, J. & Ha, Y-W. Confirmation, Disconfirmation and Information in Hypothesis Testing. Working Paper 119, Graduate School of Business, Centre for Decision Research, University of Chicago, 1986.

Kleck, R.E. Physical stigma and nonverbal cues emitted in face-to-face interaction. Human Relations, 1968, 21, 19-28.

Kruglanski, A.W. & Ajzen, I. Bias and error in human judgment. European Journal of Social Psychology, 1983, 13, 1-44.

Kruglanski, A.W., Baldwin, M.W. & Shelagh, M.J. The lay-epistemic process in attribution-making. In M. Hewstone (Ed), Attribution Theory: Social and Functional Extensions. Oxford: Basil Blackwell, 1983.

Kuhn, T.S. The Structure of Scientific Revolutions. Chicago: University of Chicago Press, 1962.

References. page 349

Langer, E.J. Rethinking the role of thought in social interaction. In J.H. Harvey, W. Ickes & R.F. Kidd (Eds.) New directions in Attribution Research (Vol. 2). Hillsdale, New Jersey: Earlbaum, 1978.

Lakatos, I. Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds), Criticism and the Growth of Knowledge. Cambridge: Cambridge University Press, 1970.

Lalljee, M. Attribution theory and the analysis of explanations. In C. Antaki (Ed.) The Psychology of Ordinary Explanations of Social Behaviour. London: Academic Press, 1981.

Levi, P. Generalisability studies in Clinical settings. British Journal of Social and Clinical Psychology, 1974, 13, 161-172.

Lipton, J.P. On the psychology of eyewitness testimony. Journal of Applied Psychology, 1977, 62, 90-95.

Loftus, E.F. Leading Questions and the eyewitness report. Cognitive Psychology, 1975, 7, 560-572.

Lord, C.G., Ross ,L. & Lepper, M.R. Biased assimilation and attitude polarisation: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 1984, 47, 1231-1243.

M

Mandler, G. Mind and Emotion. New York: Wiley, 1975.

Marshall, J., Marquis, K.H. & Oscamp, S. Effects of kind of question and atmosphere of interrogation on accuracy and completeness of testimony. Harvard Law Review, 1971, 84, 78-96.

McArthur, L.Z. The how of what and why: Some determinants and consequences of causal attribution. Journal of Personality and Social Psychology, 1972, 22, 171-193.

McGuire, W.J. The Yin and Yang of progress in social psychology: Seven Koan. Journal of Personality and Social Psychology, 1973, 26, 446-456.

Meetens, E.W., Koomann, W. Delpeut, A.P. & Hager, G.A. Effects of hypothesis and assigned task on question selection strategies. European Journal Of Social Psychology, 1984, 14, 369-378.

Merton, B.K. The Self-fulfilling Prophecy. Antioch Review, 1948, 8, 193-210.

Merton, B.K. Social Theory and Social Structure. New York: Free Press of Glencoe, 1957.

References. page 350

Miller, D.T. & Ross, M. Self-serving biases in the attribution of causality: Fact or fiction? Psychological Bulletin, 1975, 82, 213-225.

Miller, D.T. & Turnbull, W. Expectancies and interpersonal processes. Annual Review of Psychology, 1986, 37, 233-256.

Morley, I.E. & Stephenson, G.M. The Social Psychology of Bargaining. London: George, Allen & Unwin, 1977.

Moscovici, S. On social representations. In J.P. Forgas (Ed.) Social Cognition: Perspectives on everyday understanding. New York: Academic Press, 1981.

N

Nisbett, R.E. A conversation with Richard Nisbett. In E. Krupat (Ed.), Psychology is Social: Readings and Conversations in Social Psychology. Illinois: Scott, Foreman & Co., 1975.

Nisbett, R.E. & Ross, L. Human Inference: Strategies and Shortcomings of Social Judgement. New Jersey, Prentice-Hall, 1980.

Nisbett, R.E. & Wilson, T.D. Telling more than we can know: Verbal reports on mental processes. Psychological Review, 1977, 35, 231-259.

O

Orne, M.T. On the social psychology of the psychology experiment: With particular references to demand characteristics and their implications. American Psychologist, 1962, 17, 776-788.

Ornestein, A. & Phillips, W.R.E. Understanding Social Research: An Introduction. Boston: Allyn & Bacon, 1978.

P

Payne, J.W. Heuristic search processes in decision making. Advances in Consumer Research, 1975, 3, 321-327.

Peterson, C.R. & Beach, L.R. Man as an intuitive statistician. Psychological Bulletin, 1967, 68, 29-46.

Pettigrew, T.F. Extending the stereotype concept. In D.L.Hamilton (Ed.), Cognitive Processes in Stereotyping and Ingroup Behavior. New Jersey: Earlbaum, 1981.

Popper, K.R. The Logic of Scientific Descovery. New York: Harper, 1959.

References. page 351

Reis, H.T., Nezlec, J. & Wheeler, L. Physical attractiveness in social interaction. Journal of Personality and Social Psychology, 1980, 38, 604-617.

Renaud, H. & Estess, F. Life history interviews with one hundred normal American males: "Pathogenicity" of childhood. American Journal of Orthopsychiatry. 1961, 31, 796-802. (Cited in Snyder, 1981).

Riggs, J.M., Monach, E.M., Ogburn, T.A. & Pahides, S. Inducing self-perceptions: The role of social interaction. Personality and Social Psychology Bulletin, 1983, 9, 253-260.

Roisner, M. Asking silly questions. In N. Armistead (Ed), Reconstructing Social Psychology. Harmondsworth: Penguin, 1974.

Rosenfeld, H.M. Nonverbal reciprocation of approval: An experimental analysis. Journal of Experimental Social Psychology, 1967, 3, 102-111.

Rosenthal, R. Experimenter Effects in Behavioral Research. New York: Irvington, 1976.

Rosenthal, R. & Jacobson, L. Pygmalion in the classroom: Teacher expectation and Pupil's intelectual Development. New York: Holt, Rinehart & Wilson, 1968.

Rosenthal, R. & Jacobson, L.F. Teacher expectations for the disadvantaged. Scientific American, 1968.

Ross, L. The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.) Advances in Experimental Social Psychology (Vol. 10) New York: Academic Press, 1977.

Ross, L. & Andersen, C.A. Shortcomings in the attribution process: On the origin and maintainance of erronious social assessments. In D. Kahneman, P. Slovic & A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press, 1982.

Ross, L., Bierbrauer, G. & Polly, S. Attribution of educational outcomes by professional and non-professional instructors. Journal of Personality and Social Psychology, 1974, 29, 609-618.

Ross, L., Greene, D. & House, P. The "false consensus effect": An egocentric bias in social perception and attribution processes. Journal of Experimental Social Psychology, 1977, 13, 279-301.

References. page 352

Ross, L., Lepper, M.R. & Hubbard, M. Perseverence in self-perception and social perception: Biased attributional processes in the debriefing paradigm. Journal of Personality and Social Psychology, 1975, 32, 880-892.

Ross, M. Self-centred biases in attributions of responsibility: Antecedents and consequences. In E.T.Higgins, C.P. Herman & M.P.Zanna (Eds), The Ontario Symposium on personality and social psychology. Hillsdale: Earlbaum, 1981.

Ross, M. & Sicoly, F. Egocentric Biases in availability and attribution. Journal of Personality and Social Psychology, 1979, 37, 322-336.

Rothbart, M. Memory processes and social beliefs. In D.L.Hamilton (Ed.), Cognitive processes in stereotyping and ingroup behavior. New Jersey: Earlbaum, 1981.

Rothbart, M., Evans, M. & Fulero, S. Recall for confiming events: Memory processes and the maintainance of social stereotypes. Journal of Experimental Social Psychology, 1979, 15, 343-355.

Rush, M.C., Thomas, J.C. & Lord, R.G. Implicit Leadership Theory: A potential threat to the internal validity of leadership behaviour questionnaires. Organisational Behaviour and Human Performance, 1977, 21, 93-110.

S

Sackett, P.R. The interviewer as hypothesis tester: The effects of impressions of an applicant on interviewer questioning strategy. Personnel Psychology, 1982, 35, 789-804.

Schacter, S. The interaction of cognitive and physiological determinants of emotional state. In L. Berkowitz (Ed.) Advances in Experimental Social Psychology (Vol. 1). New York: Academic Press, 1964.

Schacter, S. & Singer, J.E. Cognitive, social and physiological determinants of emotional state. Psychological Review, 1962, 69, 379-399.

Seaver, W.B. Effects of naturally induced teacher expectancies. Journal of Personality and Social Psychology, 1973, 28, 333-342.

Semin, G.R. & Strack, F. The plausability of the implausable: A critique of Snyder & Swann (1978). European Journal Of Social Psychology, 1980, 10, 379-398.

References. page 353

Semin, G.R., Rosch, E & Chassein, J. A comparison of the common-sense and "scientific" condeptions of extroversion-introversion. European Journal of Social Psychology, 1981, 11, 77-86.

Shank, R.C. & Ableson, R.P. Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1977.

Shanklee, H. & Fischhoff, B. Strategies of information search in causal analysis. Memory & Cognition, 1982, 10, 520-530.

Shanklee, H. & Mims, M. Sourses of error in judging event covariations: Effects of memory demands. Journal of Experimental Psychology: Learning, Memory and Cognition, 1982, 8, 208-224.

Shaw, W.C. & Humphries, S. Influence of children's dentofacial appearance on teacher expectations. Community Dentistry and Oral Epidemiology 1982, 10, 313-319.

Short, J.A., Williams, E. & Christie, B. The Social Psychology of Telecommunications. London, Wiley International, 1976.

Singer, J.E. The use of manipulative strategies: Machiavellianism and attractiveness. Sociometry, 1964, 27, 128-150.

Smedslund, J. The concept of correlation in adults. Scandinavian Journal of Psychology, 1963, 4, 165-173.

Snyder, M. On the self-perpetuating nature of social stereotypes. In D.L.Hamilton (Ed.), Cognitive Processes in Stereotyping and Ingroup Behavior. New Jersey: Earlbaum, 1981.

Snyder, M. Seek and ye shall find: Testing hypotheses about other people. In E.T.Higgins, C.P. Herman & M.P.Zanna (Eds), The Ontario Symposium on Personality and Social Psychology. Hillsdale: Earlbaum, 1981.

Snyder, M. When Belief creates reality. In L. Berkowitz, (Ed.) Advances in Experimental Social Psychology (Vol. 18). New York, Academic Press, 1985.

Snyder, M. & Campbell, B.H. Testing of hypotheses about other people: The role of the hypothesis. Personality and Social Psychology Bulletin, 1980, 6, 421-426.

Snyder, M. & Cantor, N. Testing Hypotheses about other people: The use of historical data. Journal of Experimental Social Psychology, 1979, 15, 330-342.

References. page 354

Snyder ,M. & Gangestad, S. Hypothesis-testing processes. In J.H. Harvey, W.Ickes & R.F. Kidd (Eds.) New directions in Attribution Research (Vol. 3). Hillsdale, New Jersey: Earlbaum, 1981.

Snyder, M & Skrypnek, B.J. Testing hypotheses about the self: Assessment of job suitability: Journal of Personality, 1981, 42, 193-211.

Snyder, M. and Swann, W.B.Jr. Hypothesis-testing processes in social interaction. Journal of Personality and Social Psychology, 1978, 36, 1202-1212.

Snyder, M. & Swann, W.B. Jr. Behavioural confirmation in social interaction: From social perception to social reality. Journal of Experimental Social Psychology, 1978(b), 14, 148-162.

Snyder, M., Tanke, E.D. & Berscheid, E. Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. Journal of Personality and Social Psychology, 1977, 35, 656-666.

Snyder M. & Uranowitz, S.W. Reconstructing the past: Some cognitive consequences of person perception. Journal of Personality and Social Psychology, 1978, 36, 941-950.

Snyder,M. & White, P. Testing of hypotheses about other people: Strategies of verification and falsification. Personality and Social Psychology Bulletin, 1981, 7, 39-43.

Storms, M.D. Videotape and the attribution process: Reversing actors' and observers' point of view. Journal of Personality and Social Psychology, 1973, 27, 165-175.

Swann, W.B. Jr. The Interpersonal Nature of Self-Conceptions. Unpublished Doctoral Dissertation, University of Minnesota, 1978.

Swann, W.B. Jr., Giuliano, T. & Wegner, D. Where leading questions can lead: The power of conjecture in social interaction. Journal of Personality and Social Psychology, 1982, 42, 1025-1035.

Swann, W.B. Jr. & Hill, C.A. When our identities are mistaken: Reaffirming self-conceptions through social interactions. Journal of Personality and Social Psychology, 1982, 43, 59-66.

Swann, W.B. Jr. & Hill, C.A. The temporal stability of laboratory-induced changes in self-ratings. Unpublished Manuscript, University of Texas at Austin, 1982(b).

References. page 355

ann, W.B. Jr. & Read, S.J. Self-verification processes: How we sustain our Self-conceptions. Journal of Experimental Social Psychology, 1981, 17, 351-372.

Swann, W.B. Jr. & Snyder, M. On translating belief into action: Theories of ability and their application in an educational setting. Journal of Personality and Social Psychology, 1980, 38, 879-888.

T

Tajfel, H. Human Groups and Social Categories. Cambridge: Cambridge University Press, 1981.

Tajfel, H. & Forgas, J.P. Social categorisation: Cognitions, values and groups. In J.P. Forgas (Ed.) Social Cognition: Perspectives on Everyday Understanding. New York: Academic Press, 1981.

Tatsuoka, M.M. Multivariate Analysis. New York: Wiley, 1971.

Taylor, S.E. A categorisation approach to stereotyping. In D.L.Hamilton (Ed.), Cognitive processes in stereotyping and ingroup behavior. New Jersey: Earlbaum, 1981.

Taylor, S.E. & Fiske, S.T. Salience, attention and Attribution: Top of the head phenomena. In L. Berkowitz (Ed.) Advances in Experimental Social Psychology (Vol. 11) New York: Academic Press, 1978.

Taylor, S.E., Fiske, S.T., Etcoff, N. & Rudderman, A. The categorical and contextual bases of person memory and stereotyping. Journal of Personality and Social Psychology, 1978, 36, 778-793.

Thomas, E.A.C. & Malone, T.W. On the dynamics of two person interactions. Psychological Review, 1979, 86, 331-360.

Thomas, W.I. Situational analysis: The behaviour pattern and the situation. Reprinted in M. Janowitz (Ed.), W.I. Thomas on Social Organisation and Social Personality. Chicago: University Press, 1928/1966.

Thorndike, R.L. Review of "Pygmalion in the classroom". American Educational Research, 1968, 5, 708-711.

Topmiller, D.A. Man-Machine C^3 Simulation studies in the Air Force. In C.P. Tsokos & R.P. Thrall (Eds), Decision Information. New York: Academic Press, 1979.

Trope, Y. & Alon, E. Testing Hypotheses about the Personality of another person. Unpublished Manuscript, University of Toronto. (Cited in Trope & Bassok, 1982).

References. page 356

Trope, Y. & Bassok, M. Confirmatory and diagnostic strategies in social information gathering. Journal of Personality and Social Psychology, 1982, 43, 22-34

Trope, Y. & Bassok, M. Information-gathering strategies in hypothesis testing. Journal of Experimental Social Psychology, 1983, 19, 560-576.

Trope, Y., Bassok, M. & Alon, E. The questions lay interviewers ask. Journal of Personality, 1984, 52, 90-106.

Tversky, A. Features of similarity. Psychological Review, 1977, 84, 327-352.

Tversky, A & Kahneman, D. Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 1973, 5, 207-232.

Tversky, A. & Kahneman, D. Judgment under uncertainty: Heuristics and Biases. Science, 1974, 185, 1124-1131.

V

Vaughan, G.M. & Corballis, M.C. Beyond tests of significance: Estimating the strength of effect in selectded ANOVA designs. Psychological Bulletin, 1969, 72, 204-213.

W

Wason, P.C. On the failure to eliminate hypotheses on a conceptual task. Quarterly Journal of Experimental Psychology, 1960, 12, 129-140.

Wason, P.C. & Johnson-Laird, P.C. On the failure to eliminate hypotheses in a conceptual task. London: Batsford, 1965.

Wason, P.C. & Johnson-Laird, P.N. Psychology of Reasoning: Structure and Content. London: D.T. Batsford, 1972.

Wegner, D.M. & Vallacher, R.R. Common-sense Psychology. In J.P. Forgas (Ed.), Social Cognition: Perspectives on everyday Understanding. New York: Academic Press Inc, 1981.

Williams, E. Experimental comparisons of face-to-face and mediated communication: A Review. Psychological Bulletin, 1977, 84, 963-976.

Wilson, K.V. A distribution-free test of analysis of variance hypotheses. Psychological Review, 1956, 53, 96-101.

Winer, B.J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1971.

References. page 357

Witkins, S.L. Cognitive processes in clinical practices. Social Work, 1982, 389-395.

Wong, P.T. & Weiner, B. When people ask why questions and the heuristics of attributional research. Journal of Personality and Social Psychology, 1981, 40, 650-663.

Word, C.O., Zanna, M.P. & Cooper, J. The nonverbal mediation of self-fulfilling prophecies in interracial interaction. Journal of Experimental Social Psychology, 1974, 10, 109-120.

Z

Zajonc, R.B. Feeling and thinking: Preferences need no inferences. American Psychologist, 1980, 35, 151-175.

Zanna, M.P. & Pack, S.J. On the self-fulfilling nature of apparent sex differences in behaviour. Journal of Experimental Social Psychology, 1975, 11, 583-591.

BEST COPY

AVAILABLE

Poor text in the original thesis.
Some text bound close to the spine.
Some images distorted

Appendix 4.1

The Question list given to subjects in the experiments reported in Chapters 4 and 5

TOPIC AREAS OFTEN COVERED BY INTERVIEWERS

Below is the list of items from which you should select the 9 you want to ask during the conversation. Circle the number of each item you decide to use. Feel free to change the wording of any question if you can do so without changing its meaning.

1. Think about times when you felt lonely. What events brought on these feelings?

E
2. What events make you feel popular with people?

E
3. What activities do you really excel in?

4. In what situations do you wish you could be more outgoing?

5. Tell me about some time when you felt left out from some social group. How did you handle these feelings?

6. What kind of events make you feel like being alone?

7. What factors make it hard for you to really open up to people?

E
8. What social activities (e.g. clubs, groups.) have you been active in over the years?

E
9. What do you like about living situations in which there are always lots of people around?

E
10. What kind of situations do you seek out if you want to meet new people?

11. Describe to me a type of social situation that invariably makes you feel ill at ease and awkward. What is it about such situations that makes you uncomfortable?

12. Think about times when your shyness in social situations has made you come across as being aloof. Give me an example.

13. What things do you dislike about loud parties?

E
14. Think about times you have engaged in a lively spirited debate with someone. What are some typical things you like to debate?

E
15. In what situations are you most talkative? What is it about these situations that makes you like to talk?

16. Think about a time when you really wanted to talk to someone, but just couldn't bring yourself to initiate conversation. What types of situations are most likely to make you feel this way?

E
17. What do you like to do when you are feeling really energetic?

E
18. What would you do if you wanted to liven things up at a party?

Appendix 7.1

The Question list given to subjects in the experiment reported in Chapter 7

QUESTIONS USED IN INTERVIEWS

1 What social activities (eg. clubs, groups, etc) have you been active in over the years?

2 Would you consider yourself a shy person?

3 Do you like loud parties?

4 How long have you known your oldest friend?

5 Think about times you felt lonely. What events brought about these feelings?

6 Do you like reading books more than you like going to parties?

7 When you are put in with a group of people which you must get on with (eg. beginning a new job) do you make conversation or do you wait for others?

8 In what situations do you wish you could be more outgoing?

9 Do you live in the city or the country?

10 Do you find social gatherings awkward?

11 In what organisations do you play a leading part?

12 Do you "go out" frequently?

13 Do you have difficulty making new friends, meeting people, etc.?

14 What would you do if you wanted to liven things up at a party?

15 What sort of music do you enjoy?

16 Tell me about sometime when you felt left out from some social group. How did you handle these feelings?

17 In a party where you knew few people, would you try to get to know someone or would you sit in a corner and wait for them to come to you?

18 Would you describe yourself as sociable?

19 Do you prefer to go to wild parties or sedate dinner parties?

20 Do you often have long serious discussions with close friends?

BI
21 What factors make it really hard for you to open up to people?

22 Do you consider yourself to have had a upbringing with a strong emphasis on discipline and conformity?

23 In what situations are you most talkative? What is it about these situations that makes you like to talk?

N FEEL
24 When in a social situation with other people, do you relaxed and confident or on-edge and uncomfortable?

E I
25 What things do you dislike about loud parties?

26 Are you nervous about interviews where you have to make a good impression?

27 What do you like about living in situations where there are a lot of people around you?

E
28 Do you like to feel that other people's attention is focused on you?

29 What is your opinion of marriage?

E
30 Do you adjust to new environments quickly, making new friends etc.?

BI
31 Describe to me a type of social situation that invariably makes you feel ill at ease and awkward. What is it about these situations that makes you feel uncomfortable?

N
32 How do you find that others view you in regard to your temperament etc?

33 What do you like to do when you are feeling really energetic?

34 What do you think of Margaret Thatcher?

E
35 Have you ever been accused of talking too much or dominating conversations?

36 Think about times when your shyness in social situations has made you come across as being aloof. Give me an example.

37 How do you see the future of the world?

38 Do unfamiliar situations and occurrances make you feel uneasy?

39 As far as you can tell does your entheusiasm rub off on those around you?

N
40 Do you prefer to spend free time in the company of others, alone, or a mixture of the two?

41 In company do you enjoy bragging about your exploits?

42 Would you say that you enjoyed spending time on your own?

43 Do you believe in your own destiny, or do you think that others are controlling it?

44 Think about times when you have engaged in a lively and spirited debate with people. What are some typical things you like to debate?

45 Would you like to become famous?

46 Do you find it hard to make conversation with strangers?

47 Do you prefer having many aquaintances rather than few but special friends?

48 Think about a time when you really wanted to talk to someone but just couldn't bring yourself to initiate conversation. What types of situations are most likely to make you feel this way?