Quarterly Journal of Experimental Psychology (1977) 29, 85–95

CONFIRMATION BIAS IN A SIMULATED RESEARCH ENVIRONMENT: AN EXPERIMENTAL STUDY OF SCIENTIFIC INFERENCE

CLIFFORD R. MYNATT, MICHAEL E. DOHERTY AND RYAN D. TWENEY

Psychology Department, Bowling Green State University, Bowling Green, Ohio 43403, *U.S.A.*

Numerous authors (e.g., Popper, 1959) argue that scientists should try to *falsify* rather than *confirm* theories. However, recent empirical work (Wason and Johnson-Laird, 1972) suggests the existence of a confirmation bias, at least on abstract problems. Using a more realistic, computer controlled environment modeled after a real research setting, subjects in this study first formulated hypotheses about the laws governing events occurring in the environment. They then chose between pairs of environments in which they could: (1) make observations which would probably confirm these hypotheses, or (2) test alternative hypotheses. Strong evidence for a confirmation bias involving failure to choose environments allowing tests of alternative hypotheses was found. However, when subjects did obtain explicit falsifying information, they used this information to reject incorrect hypotheses.

Introduction

The present study investigates the behaviour of people engaged in testing scientific hypotheses in a dynamic, artificial environment. Such hypothesis testing can be regarded from two distinct, though overlapping, perspectives. First, the relationship between the universal statements of science (general laws and theories) and the empirical data relevant to these statements can be examined from a purely logical point of view. This approach, which can be characterized as formal, analytic and, in a sense, normative, has been common among philosophers of science. A second approach is to consider scientific inference as a psychological problem involving decisions about the relationship between a scientific theory and a particular body of evidence. This approach, which can be characterized as empirical, synthetic and descriptive, is rare among philosophers and, surprisingly, among behavioural scientists as well. Exceptions do exist (e.g. Kuhn, 1970; Mitroff, 1974), but most analyses of science have been formal rather than empirical.

The question of the relationship between evidence and the truth status of a scientific theory was first raised in full force by Hume (1955, originally published

in 1748). Hume asserted that it is not possible to justify belief in the truth of a general scientific law or theory no matter how much evidence is produced to support it. His argument is based on the fact that while scientific propositions are universal statements which have an infinite number of predicted consequences (e.g. $F = ma$, which refers to *all* force, mass, and acceleration at *all* times and *all* places), the evidence for such statements is always finite (e.g. those particular instances of force, mass, and acceleration which have actually been observed). Thus, Hume concluded, inductive inference is not a justifiable procedure.

Numerous counter-arguments to Hume's skeptical conclusions have been proposed (see Salmon, 1967), one of the more widely accepted being Popper's (1959, 1962) falsification position. Given a conditional statement of the form "If P then Q", showing that Q is true does not establish either the truth or falsity of P. However, if Q is false, then P must be false. Since nearly all scientific propositions are of this form, where P is a general theory or law and Q a predicted event, Popper argues that the only justifiable inference from data about such propositions is whether or not they have been falsified. He concludes that scientists should not attempt to confirm hypotheses, but rather should conduct research so as to maximize the likelihood of disconfirmation.

A related position, based on pragmatic and historical rather than logical considerations, is Platt's (1964) "strong inference" strategy. This consists of first generating multiple hypotheses relevant to a particular phenomenon and then performing experimental tests to eliminate (i.e. falsify) as many of these as possible. The procedure involves a cycle of multiple hypotheses, experimental elimination of hypotheses, new multiple hypotheses, etc. Platt urges that, above all, the scientist should not focus upon the confirmation of a single, favorite hypothesis and thereby become wedded to an incorrect theory.

Popper's and Platt's prescriptions for scientific inquiry differ in one important respect. Popper exhorts the scientist to direct his attention to a particular kind of data search, i.e. data inconsistent with the particular hypothesis under test. Platt accepts the utility of falsification, but stresses that it should be done in the context of multiple alternative hypotheses.

Both positions raise important questions about the actual behaviour of scientists (and, indeed, people in general) which cannot be answered adequately by either logical or historical analysis. One question is whether disconfirmatory evidence is sought or recognized; a second is whether alternative hypotheses are sought or examined. Anecdotal evidence seems to suggest that the answer to both questions may be "no". Examples abound of scientists clinging to pet theories and refusing to seek alternatives in the face of large amounts of contradictory data (see Kuhn, 1970). Objective evidence, however, is scant.

Wason (1968a) has conducted several experiments on inferential reasoning in which subjects were given conditional rules of the form "If P then Q", where P was a statement about one side of a stimulus card and Q a statement about the other side. Four stimulus cards, corresponding to P, not-P, Q, and not-Q were provided. The subjects' task was to indicate those cards—and only those cards—which had to be turned over in order to determine if the rule was true or false. Most subjects chose only P, or P and Q. The only cards which can falsify the rule,

however, are P and not-Q. Since the not-Q card is almost never selected, the results indicate a strong tendency to seek confirmatory rather than disconfirmatory evidence. This bias for selecting confirmatory evidence has proved remarkably difficult to eradicate (see Wason and Johnson-Laird, 1972, pp. 171–201).

In another set of experiments, Wason (1960, 1968*b*, 1971) also found evidence of failure to consider alternative hypotheses. Subjects were given the task of recovering an experimenter defined rule for generating numerical sequences. The correct rule was a very general one and, consequently, many incorrect specific rules could generate sequences which were compatible with the correct rule. Most subjects produced a few sequences based upon a single, specific rule, received positive feedback, and announced mistakenly that they had discovered the correct rule. With some notable exceptions, what subjects did not do was to generate and eliminate alternative rules in a systematic fashion. Somewhat similar results have been reported by Miller (1967).

Finally, Mitroff (1974), in a large-scale non-experimental study of NASA scientists, reports that a strong confirmation bias existed among many members of this group. He cites numerous examples of these scientists' verbalizations of their own and other scientists' obduracy in the face of data as evidence for this conclusion.

All the above evidence suggests that a bias in favour of confirmatory evidence may be a general, trans-situational characteristic of human reasoning. This bias may be expressed either as a failure to seek or utilize data which are inconsistent with a single hypothesis under test, contrary to Popper's falsification strategy, or it may be expressed as a failure to seek or utilize evidence for alternative hypotheses, contrary to Platt's strong inference strategy.

Unfortunately, there has been little experimental research on reasoning and inference processes in settings which resemble real world situations such as those in which scientific inquiry occurs. The Wason research, which appears to us to be dealing most closely with the type of reasoning involved in scientific inference, has utilized relatively content-free tasks which lack much of the concrete quality of science as it occurs in the real world. Indeed, in at least one situation in which subjects were given a realistic, non-abstract task (Johnson-Laird, Legrenzi and Legrenzi, 1972), they seemed quite capable of seeking and utilizing falsifying information. It is not presently known, however, what specific task requirements are necessary to produce this result.

The purpose of the present study is to investigate inference behaviour in a setting designed to resemble the conditions under which actual science is done. It poses two questions. First: Do subjects tend to select situations for testing their hypotheses which allow only confirmatory observations rather than selecting situations which allow alternative hypotheses to be tested? Second: Do subjects who obtain direct falsifying evidence change hypotheses?

A fairly complex, dynamic environment was employed in which subjects were able to test concrete hypotheses by selecting various experiments to perform. Some of these experiments could, given a particular hypothesis, produce only confirmatory evidence for that hypothesis. Other experiments allowed the subject to test alternative hypotheses. One allowed direct falsification.

Method

The simulated scientific environment was designed as a controlled setting in which the behaviour of subjects could be closely monitored and which would mirror the qualities of actual research situations. The following general characteristics were deemed necessary for any such system:

(1) The environment should have an "object quality". That is, events and objects should have a perceptual "reality" similar to events and objects in the real world.

(2) The environment should be lawful. A few simple equations should describe fully the relationships among the elements of the system.

(3) The environment should appear complex. The simplicity of the underlying laws should be apparent only after a large amount of careful observation and experimentation.

(4) The environment should be dynamic and interactive. Subjects should be able to make observations, plan and carry out experiments, etc.

(5) The attempt to understand the environment should be interesting and challenging.

These characteristics were implemented by using an on-line computer which allowed the artificial research environment to be visually displayed to subjects and to be manipulated by them in an interactive mode.

Subjects

Forty-five undergraduates drawn from the Introductory Psychology subject pool at Bowling Green State University served as subjects.

Apparatus

The computer was a Data General NOVA 1220 mini-computer interfaced to an Owens-Illinois Digivue plasma display screen. A keyboard immediately below the screen was used for subject/machine interaction. The screen consists of a 512 × 512 grid. Each of the 262 144 points formed by this grid is separately addressable and, when addressed, glows with an easily visible orange light. Overall display size is a 21·6 cm square.

The research environment was programmed in GBASIC, a version of BASIC especially written for the Digivue/NOVA combination (Fulton, 1974). The program allowed stationary figures with one of three different shapes (triangles, squares or discs) to be displayed. Each stationary figure had one of two different brightness levels: all points enclosed by the figure lit or half of the points enclosed by the figure lit. These levels will be identified subsequently as the ratio of the number of points lit to total points, that is, 1·0 and 0·5, respectively. A particular arrangement of figures will be referred to as a "screen". Each such screen corresponds to one view of the total imaginable universe defined by the laws of the system.

A programmed keyboard command could be used by subjects to cause a small lighted dot or "particle" to move by small steps from a fixed position in the upper left corner of each screen at a speed of approximately 0·6 cm/s toward any point on the screen. A circular, non-visible boundary extended 4·2 cm beyond the geometric centers of some of the figures. Whenever a particle encountered a boundary, its motion ceased. Only figures with brightness ratios of 0·5 had such boundaries. All other aspects of the stationary figures, such as size, shape, location, etc., had no effect on particle motion and, hence, were irrelevent cues.

Procedure

Subjects were given printed instructions on how to use the keyboard, then shown the first screen which contained a 0·5 triangle, a 1·0 square and a 1·0 disc. They could fire as many particles as they wished at any part of the screen, following which they were asked to write down an hypothesis which would account for the motion of the particles.

They were then shown a second screen on which were several 1·0 squares and discs, a 0·5 triangle, and a 1·0 triangle in close proximity to a 0·5 disc (see Fig. 1). The last two

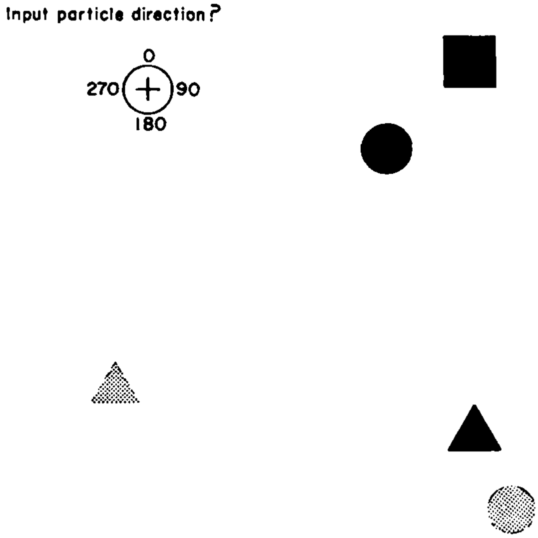

FIGURE 1. Second screen. Textured features are those having 0·5 brightness levels. Particles were fired from the grid in the upper left of the screen.

features were so close together that the triangle was completely inside the boundary of the disc and were arranged so that the disc was behind the triangle, relative to the location form which particles were fired. Subjects again fired as many particles as they wished and were instructed to write down an hypothesis which would account for the motion of the particles on both of the first two screens. They were told that this hypothesis could be the same as their first one, or that it could be different if evidence from the second screen had changed their minds. This hypothesis will be identified subsequently as the subjects' "initial hypothesis".

The purpose of the first two screens was to make probable the occurrence of hypotheses focussing on triangles. Shape and brightness were confounded on both screens, and we assumed that shape would be the more salient cue.

Subjects were then randomly assigned to one of three instructional treatments:

(1) *Instructions to confirm*

Subjects were given written instructions which stated that the basic job of a scientist was to confirm theories and hypotheses. They were given an historical example of such a confirmation and told to try to confirm their hypotheses about particle motion.

(2) *Instructions to disconfirm*

These were identical to the confirmation instructions except that subjects were told that the basic job of a scientist was to disprove or disconfirm theories and hypotheses. An historical example of scientific disconfirmation was given and subjects were instructed to attempt to disconfirm their hypotheses.

(3) *Instructions to test*

These were identical to the preceding instructions except that subjects were simply told that the job of a scientist was to "test theories and hypotheses", and that they were to test their own hypotheses about particle motion. They were given no further examples or instructions.

Dependent variables

Paired screen choices

After assignment to an instructional treatment, all subjects were shown 10 pairs of photographs of screens and asked to choose the one member of each pair on which they would prefer to fire additional particles to obtain more evidence concerning their hypothesis. These photographs were juxtaposed in a ring binder, one pair on a page. Six of the pairs were designed so that, given a triangle hypothesis, one of the screens could produce only confirmatory evidence (see Table 1). For example, on pair 9, screen A contained a single 0·5 triangle and screen B contained a single 0·5 disc. The subjects could choose a screen (A) which contained a feature similar to features which had apparently stopped particles on the first two screens; or they could choose a screen (B) with a feature which had not been previously encountered in isolation. Subjects who were seeking confirmatory evidence for the hypothesis "Triangles stop the particle" should, therefore, choose (A) over (B). The other four pairs were designed so that neither screen would be more likely than the other to produce confirmatory evidence.

TABLE I
Paired screen descriptions

Pair No.	Screen A	Screen B
1	†0·5 disc; 1·0 square	*0·5 triangle; 1·0 square
2	*0·5 triangle	†0·5 square
3	1·0 square; 1·0 disc	1·0 square; 1·0 square
4	†0·5 disc; 0·5 disc	*0·5 disc; 0·5 triangle
5	0·5 triangle	‡1·0 triangle
6	1·0 square	1·0 disc
7	*1·0 triangle; 0·5 square	†1·0 disc; 0·5 square
8	1·0 square	1·0 disc; 1·0 disc
9	*0·5 triangle	†0·5 disc
10	†1·0 disc; 0·5 disc	*0·5 triangle; 1·0 disc

Note. For all screens, the particle firing grid was in the upper left corner as in Fig. 1. All stationary figures were positioned in the lower middle or the lower right corner of the screens. When two figures were present, the distance between the centers of the figures was approximately 2 cm (e.g. the disc and triangle in the lower right corner of the screen shown in Fig. 1).

*Confirmatory choices for subjects with triangle hypotheses.

†A particle fired at the figures on this screen would stop, providing evidence for an alternative to a triangle hypothesis.

‡A particle fired at the triangle on this screen would not stop, logically disconfirming a triangle hypothesis.

Subjects who were attempting to test alternatives to a triangle hypothesis should show a very different set of choice responses. On pair 9, for example, such a subject should not select screen A (a single 0·5 triangle). Rather, he should select screen B (a single 0·5 disc), since he can on this screen test the hypothesis that some characteristic of the features other than triangularity affects particle motion (e.g. brightness ratio).

One screen, 5-B can provide absolute, unambiguous disconfirmation of a triangle hypotheses. On this screen, a particle would travel completely through the 1·0 triangle.

Free responses

After the paired screen tasks, subjects were given the opportunity to fire particles on either chosen or non-chosen screens, but only one member of each pair (a restriction some

subjects ignored). They were instructed to record what happened on each firing; whether or not their hypotheses changed and, if so, why; and, after firing as many particles as they wished on as many of the 10 screens as they wished, to write down their final hypotheses. All subject input to the computer was automatically recorded. Thus, it was possible to determine, for example, whether a subject had ever observed the particle stopping on one of the screens not containing the triangles (e.g., 9-B).

Results

Subjects were first categorized by their initial hypotheses. Three categories were used: (a) Triangle, which included hypotheses with at least some mention of triangularity but no mention of brightness; (b) Brightness, which included hypotheses with any mention at all of brightness; and (c) Other. Across conditions, 20 subjects had initial triangle hypotheses; 12 subjects had initial brightness hypotheses; and 13 subjects had initial hypotheses which were classified as "other". Of the 20 subjects classified as having initial triangle hypotheses, none mentioned brightness and 15 mentioned nothing other than triangularity. These 20 subjects are of primary interest since the paired screen task was designed for subjects with a triangle hypothesis.

Paired screen data

The paired screen choice data for the 20 subjects with initial triangle hypotheses are shown in Table II. The choice data were collapsed across the three conditions,

TABLE II

Paired screen choice data for subjects with initial triangle hypotheses (N = 20)

Condition	
Test	$n = 7$ 0·71
Disconfirm	$n = 6$ 0·70
Confirm	$n = 7$ 0·71
Overall mean proportion = 0·71	

Note. Cell entries are number of subjects (n) in each condition who had an initial triangle hypotheses, and the mean proportion of confirmatory choices for these subjects.

producing an overall mean proportion of confirmatory choices of 0·71. This mean proportion differed significantly from an expected mean proportion of 0·50 under a null hypothesis of random choice ($t = 4{\cdot}31$, $df = 19$, $P < 0{\cdot}01$). That this effect was due to the subjects' initial hypothesis, and not to some artifact of the choice procedure, is demonstrated by the fact that the comparable proportions for subjects with initial brightness hypotheses (0·44 across conditions) and other hypotheses (0·47 across conditions) did not differ from an expected mean proportion of 0·50. The tendency to choose confirmatory screens was present from the very first pair of screens. Fifteen of the 20 subjects chose screen B (a confirmatory choice) on pair 1.

Free response data

The free response data of the 20 subjects with initial triangle hypotheses were examined with three questions in mind:

(1) Did subjects obtain evidence logically falsifying a triangle hypothesis by firing a particle on screen 5-B at the 1·0 triangle?

(2) Did subjects make observations which would unambiguously support alternative hypotheses by firing a particle on a screen where a figure other than a triangle stopped the particle motion?

(3) If subjects obtained evidence of the type described in (1) and (2), what changes, if any, did they make in their incorrect initial hypotheses?

Final hypotheses were categorized as being: (a) correct or partially correct (i.e. brightness clearly identified as the only variable influencing particle motion or some mention of brightness but with irrelevant variables added); or (b) incorrect. Any of the 20 subjects whose final hypothesis is in category (a) has, of course, changed his hypothesis. The relevant data are summarized in Table III.

TABLE III

Free response data for subjects with initial triangle hypotheses (N = 20)

Free response choices	Correct or partially correct	Incorrect
No Fal No Alt	2	2
No Fal Alt	2	3
Fal No Alt	0	0
Fal Alt	10	1

Note. *Fal* means that the subject fired a particle at the triangle on screen 5-B and has therefore received evidence falsifying a triangle hypothesis, *Alt* means that the subject fired a particle at a 0·5 brightness, non-triangular feature and has therefore received evidence supportive of an alternative to a triangle hypothesis.

Of the 11 subjects who made logically falsifying observations, 10 had final hypotheses which were correct or partially correct. Of the nine who did not make a falsifying observation, only four achieved a correct or partially correct final hypothesis. These differences are significant using a χ^2 test ($\chi^2 = 5{\cdot}08$, $df = 1$, $P < 0{\cdot}05$).

Final hypothesis data

All subjects, regardless of initial hypothesis, were categorized by their final hypotheses. The same two categories (correct or partially correct, and incorrect) were again used. These data are shown in Table IV. A χ^2 test of the frequencies

TABLE IV

Final hypotheses data for all subjects (N = 45)

Condition	Correct or partially correct	Incorrect
Test	11	4
Disconfirm	11	4
Confirm	6	9
Overall	28	17

Cell entries are number of subjects in each category.

in the confirm and disconfirm conditions approached, but did not reach, significance ($\chi^2 = 3{\cdot}39$, $df = 1$, $P < 0{\cdot}10$).

Initial vs. final hypothesis data

Finally, all subjects were categorized by both initial and final hypotheses. The same categories as above were used. There were no significant differences across instructional treatments and the data were collapsed across treatments. These data are shown in Table V. A Chi-square test of the frequencies in Table V was significant ($\chi^2 = 13{\cdot}42$, $df = 2$, $P < 0{\cdot}01$).

TABLE V

Initial and final hypotheses for all subjects (N = 45)

Initial hypothesis	Final hypothesis Correct or partially correct	Incorrect
Brightness	11	1
Triangle	14	6
Other	3	10
Overall	28	17

Cell entries are number of subjects in each category across all three conditions.

Discussion

Two major conclusions can be drawn from these results. First, there is substantial evidence from the paired screen choice data that subjects failed to consider alternative hypotheses. Subjects who started with triangle hypotheses, regardless of which condition they were in, chose at a much higher than chance rate screens which could only confirm such hypotheses. Thus, they did not, in general, choose screens which would allow them to test alternatives to their initial hypotheses (e.g. by choosing screens with a square or disc). These results, which are remarkably similar to those found by Wason (1960, 1968*b*, 1971) suggest that confirmation bias of this sort may be a general cognitive process which is not limited to abstract tasks. Wason's subjects generally failed to test alternatives when

attempting to recover a numerical rule, just as our subjects failed to test alternatives when trying to discover the laws of our artificial universe.

Second, while the paired screen choice data provide strong support for a bias involving failure to consider alternative hypotheses, the free response data indicated that subjects could use falsifying data once they got it. Among those subjects who had an initial triangle hypothesis, nearly all (91 %) who obtained unambiguous falsifying evidence for this hypothesis changed to a correct or partially correct hypothesis. Less than half (44 %) of such subjects who did not obtain falsifying evidence reached a correct or partially correct solution. Thus, if confronted with unambiguous falsification, subjects appeared to appreciate its impact. It is interesting to note in this context that 92 % of the subjects who mentioned brightness in their initial hypotheses (i.e. had approximately correct initial hypotheses) had at least a partially correct final hypothesis. Almost exactly the same percentage, 91 %, of subjects who had initial triangle hypotheses but got disconfirming evidence had at least a partially correct final hypothesis.

These two conclusions are reinforced by the results of a subsequent replication experiment. Thirty subjects were run under test condition instructions and produced data almost identical to those of the original study. For initial triangle hypothesis subjects ($n = 18$), the mean proportion of confirmation choices was 0·70. Among these subjects, those who received falsifying evidence were again significantly more likely to achieve a correct final hypothesis than were those who did not make a falsifying observation.

These findings have several implications with respect to both philosophical analysis of falsification and prior research on confirmation bias. Wason (1968*a*) suggests that his data from the four-card selection task indicate that people have difficulty understanding the logic of falsification. Our subjects experienced no such difficulty. When confronted with unambiguous falsifying evidence, they utilized it in precisely the correct way—by rejecting their incorrect hypotheses, just as Popper says they should. Our subjects did not, on the other hand, appear to look for and test alternative hypotheses. Thus, at least in the absence of explicit training or experience, people may not utilize anything like the multiple hypothesis approach advocated by Platt. Such a failure to look for alternative hypotheses becomes even more interesting when the relationship in the present study between initial and final hypotheses is examined. Of those subjects whose initial hypothesis at least mentioned brightness as a possible variable, nearly all (92 %) arrived at a correct or paritally correct final hypothesis. Only about half (52 %) of the subjects whose initial hypothesis did not mention brightness arrived at a correct or partially correct solution. This indicates that the hypothesis generation stage in an inference process may be of critical importance. If an initial hypothesis is totally incorrect and misleading, and if alternatives are not considered, then arriving at a correct hypothesis may be very difficult. On the other hand, the effect of confirmation bias may not be so disadvantageous if the initial hypothesis is at least partly correct.

Platt's strong inference model should be considered in the light of the results of this experiment, although much more research must be done before too much is claimed. Three aspects of the results are relevant to Platt's approach. The

first is the relative success of subjects who attended to the correct dimension in the hypothesis generation stage. The probability of doing so should be considerably enhanced if multiple hypotheses are explicitly tested. The second is the failure of subjects to choose test screens allowing the test of alternative hypotheses. The explicit formulation of multiple hypotheses should preclude the operation of this source of confirmation bias. Third, subjects in this study used falsifying evidence when they obtained it. Thus, Platt's model, which assumes some ability to falsify, is at least not demanding the impossible, even of naive subjects. Whether subjects can readily design experiments to falsify, which is what Platt prescribes, is unclear. However, the weakness of the instructional manipulation, and some yet unpublished, subsequent research on the topic, suggest that such behaviour is very difficult to elicit.

References

FULTON, D. L. (1974). A plasma-panel interactive graphic system. *Proceedings of the Society for Information Display*, **15**, 74–80.

HUME, D. (1955). *An Inquiry Concerning Human Understanding*. New York: Bobbs-Merrill.

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

KUHN, T. S. (1970). *The Structure of Scientific Revolutions*, 2nd ed. Chicago: University of Chicago Press.

MILLER, G. A. (1967). Project Grammarama. In *The Psychology of Communication*. New York: Basic Books.

MITROFF, I. I. (1974). *The Subjective Side of Science*. Amsterdam: Elsevier Scientific Publishing Co.

PLATT, J. R. (1964). Strong inference. *Science*, **146**, 347–53.

POPPER, K. R. (1959). *The Logic of Scientific Discovery*. New York: Basic Books.

POPPER, K. R. (1962). *Conjectures and Refutations*. New York: Basic Books.

SALMON, W. (1967). *The Foundations of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.

WASON, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, **12**, 129–40.

WASON, P. C. (1968a). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, **20**, 273–81.

WASON, P. C. (1968b). On the failure to eliminate hypotheses in a conceptual task: a second look. In WASON, P. C. and JOHNSON-LAIRD, P. N. (Eds), *Thinking and Reasoning*. Baltimore: Penguin Books.

WASON, P. C. (1971). Problem solving and reasoning. In SUMMERFIELD, A. (Ed.) *Cognitive Psychology*, British Medical Bulletin, **27**, 206–10.

WASON, P. C. and JOHNSON-LAIRD, P. N. (1972). *Psychology of Reasoning: Structure and Content*. Cambridge, MA: Harvard University Press.

Received 30 *January* 1976