WHEN FALSIFICATION FAILS

Hugh Garavan Michael E. Doherty Clifford R. Mynatt

Bowling Green State University
Bowling Green, OH 43403, USA

Mail correspondence to:
Hugh Garavan, Ph.D.
Department of Psychology,
Trinity College,
Dublin 2,
Ireland

Ph: +353-1-608-3910
Fax: +353-1-671-2006
E-mail: Hugh.Garavan@tcd.ie
http://www.tcd.ie/Psychology/Hugh_Garavan/

This study investigated the effectiveness of a falsification logic at early and late stages of the hypothesis testing process. The subject's task was to discover the "laws of motion" in a computerized Artificial Universe. Twenty science students, after detailed instructions on hypothesis testing, worked independently on the task in the presence of the experimenter. Subjects who were instructed to generate multiple hypotheses and follow a falsification logic did significantly worse than those receiving no such instructions. The manipulation of the early and late knowledge stages had no effect, apparently because few subjects attended to the information intended to put them into an advanced state of knowledge. A combination of quantitative and qualitative analyses suggested the following generalizations. 1. While falsifiability may be a useful criterion for demarcating science from non-science, the deliberate attempt to use falsification may be counterproductive in complex environments. 2. The use of a falsification logic may simply be beyond the cognitive capacity of people in complex environments; it is very difficult to try to prove something false that one believes true. 3. The use of metaphors and analogies, largely precluded from functionality in most laboratory reasoning tasks, appears spontaneous and useful in richer environments. 4. In complex environments, it is evident that a richer classification scheme for the types of hypothesis tests than currently accepted may be required.

INTRODUCTION

What rules should scientists follow to pursue the investigation of the world around them? This is a question that has concerned scientists and philosophers alike, and is the topic of the present paper. Popper (1962) argued that it is the very fact that scientific hypotheses can be falsified that gives science its unique status as a means to knowledge. He noted that it is easy to obtain confirmations or verifications of a theory, but that such confirmations are no guarantee that the theory is true. According to Popper, it is in the disproving of a theory or a hypothesis that progress is made; hence the only good tests of a theory are those that try to disprove it. Scientists must, he advocated, strive to disprove what they consider true. Assuming that there are competing theories available, the implementation of the strategy of falsification should lead science to progress in a manner analogous to natural selection.

Popper accepted that positive successes,that is, confirmations of a theory, are essential to scientific progress, but only when they result from unsuccessful attempts to falsify that theory. Without successes, he argued, we cannot appreciate the significance of successful and enlightening refutations. Essentially then, verifications and refutations must go hand in hand; "...both kinds of successes are essential: success in refuting our theories, and success on the part of some of our theories in resisting at least some of our most determined attempts to refute them" (1962, p. 245).

In some other approaches to the philosophy of science (Lakatos, 1970; 1978; Quine, 1961), falsifying evidence is attributed less importance. A more sophisticated falsificationism stresses the relative falsifiability of competing theories as well as the importance of confirming bold conjectures (Chalmers, 1982, Ch.5). Nonetheless, the legacy of Popper remains strong and falsifiablity is often held as a standard for demarcating science from non-science. A multiple hypothesis and disconfirmationist approach to the pursuit of science has been argued by some scientists (Eccles, 1981; Platt, 1964; Sekuler, 1976) to be the antidote to the possibly maladaptive "confirmation biases" that have been revealed in people's thinking (Gilovich, 1991; Kahneman, Slovic & Tversky, 1982). Despite its inadequacies as a valid description of the everyday practice of science, the prescriptive aspect of falsificationism survives.

Confirmation Bias in Science

In contrast to the imperative to falsify, research on hypothesis testing, both on scientists and (more commonly) on subjects under experimental conditions, has shown an overwhelming tendency to seek data consistent with the hypothesis that one is testing. This has been labeled confirmation bias (Tweney, Doherty & Mynatt, 1981; Wason, 1960), and more recently, a positive test strategy (Klayman & Ha, 1987) or positivity bias (Evans, 1989). This tendency can affect both the information that one selects, as in Wason's four-card task (Wason & Johnson-Laird, 1972), and one's interpretation of new information that bears on the truth or falsity of one's original hypothesis, as in the pseudodiagnosticity research (Doherty, Mynatt, Tweney, & Schiavo, 1979; Doherty, Chadwick, Garavan, Barr, & Mynatt, 1996). If a scientist looks only for information that supports his or her own theory then that scientist may never find evidence that will disprove the theory, or, indeed, support an alternative one. If results are only interpreted within the framework of a particular theory then one will see not refutations and disproofs, but exceptions and anomalies (Kuhn, 1962). The tendency to confirm is pervasive. Not only has it been reported in many different experimental tasks, some of which will be considered in more detail below, but it has also been reported among active scientists. Mitroff (1974) conducted interviews with 40 NASA scientists and found that not only were they deeply committed to their own theories but that they considered this a commendable trait.

A number of other investigations have also been carried out with scientists to see the extent to which a confirmation bias is present. One, by Mahoney & DeMonbreun (1977), compared the problem-solving skills of 30 Ph.D scientists with those of 15 Protestant ministers. Their reasoning
skills did not differ. Another study (Kern, Mirels and Hinshaw, 1983) assessed scientists' understanding of basic principles of formal deductive logic. They presented 72 scientists with problems dealing with, for example, the logical validity of modus ponens and modus tollens. "Perhaps the most striking finding . . . was the failure of nearly half of the scientists to recognize the logical validity of modus tollens. This is particularly noteworthy given the normative view that disconfirmation should play a major role in theory and hypothesis testing." (p.142).

Laboratory Simulations of Science

Wason's 2-4-6 task

Wason (1960, 1968) developed what he considered a task that "simulates a miniature scientific problem" (1960, p. 139). In this task the experimenter has a rule in mind that determines whether or not a triple (any set of three numbers) is included in a target set. The experimenter starts by giving an example of a triple that conforms to the rule, namely "2-4-6." Subjects then propose triples of their own and the experimenter indicates whether or not each is a member of the target set. Subjects typically proceed by generating a hypothesis (commonly, "numbers increasing by two" or "even numbers"), test that hypothesis several times and then, when they feel "highly confident" that they know the experimenter's rule, propose it, only to have it rejected. The correct rule in the standard version of the task is "any three numbers increasing in value." Wason reported a common and maladaptive (in this task) tendency towards a confirmation bias; a majority of the subjects terminated their hypothesis testing highly confident of an incorrect rule. This finding has been replicated many times, even with samples of engineers, scientists, and statisticians (e.g., Einhorn & Hogarth, 1978; Mahoney & DeMonbreun, 1977). In his study, Wason found that successful solution was associated with greater use of "eliminative tests." The general finding though, has been that subjects rarely attempt to falsify the single hypothesis that they have generated, and rarely propose multiple hypotheses.

There have been a number of variants on the standard 2-4-6 task. Experiments I and II of Tweney, Doherty, Worner, Pliske, Mynatt, Gross, and Arkkelin (1980) found that instructions to disconfirm increased the number of disconfirmatory triples, but did not result in more success in solving the problem. Nor did teaching subjects to use multiple alternative hypotheses (Experiment III). In Experiment IV, while the formal structure of the task remained the same, its psychological structure was altered by asking the subjects to discover two interrelated rules. It became a two-category classification task such that instead of "right" triples and "wrong" triples there were now triples that were either DAX or MED. The effects on performance were dramatic; 60% of the subjects correctly determined the DAX rule on their first rule announcement. Apparently "negative information became relevant information, instead of something to be ignored" (p.122). Many subjects adopted a strategy of alternating their testing between the DAX and MED rule. This approach proved very effective, thus reaffirming the value of giving consideration to more than just one hypothesis.

Subjects in the DAX-MED condition proved most successful even though they made fewer disconfirmatory tests than subjects in Experiments I and II, proceeding rather with confirmatory tests of two separate hypotheses. A similar finding was also noted by Gorman, Stafford and Gorman (1987), who used a more difficult version of the 2-4-6 task. They reported that 88% of the subjects solved the task on the DAX and MED condition as opposed to just 21% on the control condition. They further reported that "DAX-MED subjects tended to search for positive instances of the MED rule, which, in turn, forced them to test the limits of the DAX rule" (p.1).

Artificial Universe tasks

Other investigators have employed more complex task environments, in the belief that the focus on falsification may have different effects in such environments than in many of the typical logical reasoning tasks. Mynatt, Doherty & Tweney (1978) presented subjects with a computer environment, a nine-screen Artificial Universe, depicted in Figure 1. This Universe contained 27 objects of different brightness levels, shapes and sizes. The subject’s task was to discover the rules that determined the motion of particles within this Universe. To this end subjects were allowed to perform experiments which entailed firing a particle from any location in any screen and in any direction. Figure 1 shows the boundaries that surrounded two-thirds of the objects. These boundaries deflected particles but were not visible to subjects. As with the 2-4-6 task, confirmatory strategies were the rule, even though half of the subjects had been given extensive instructions in strong inference (Platt, 1964) and in the logical value of falsification. In fact, those few subjects who readily abandoned hypotheses in the face of disconfirming evidence were further from discovering the laws of the Universe at the conclusion of the study than they had been at the start.

This result suggested "that the use of disconfirmation may not be a universally effective inference technique. This may be especially true in the early stages of a complex inference task. It may not be psychologically possible to establish useful inductive regularities and to generate good alternative hypotheses at the same time. In fact, it may not be possible to establish such inductive regularities without some theory, even an incorrect one" (Mynatt et al., 1978, p.405). These authors suggested that "... traditional philosophical analysis of falsification may not have recognized some important psychological realities," and that "...empirical investigation into the roles of confirmation and disconfirmation at different stages during inference processes may be richly rewarding" (p.405).

Writing later, Tweney (1989) stated that this work "with 'artificial Universe' tasks led us to the formulation of the 'early-late' notion of heuristic organization, namely that seeking confirmation is optimal early in the hypothesis testing process, whereas a shift to seeking disconfirmation is optimal later in the process" (p.349). This early-late hypothesis is of central importance to the present study, but the terms early and late refer to the person's degree of task relevant knowledge rather than to the time one has spent on the task. These are, of course, usually related.

Insert Figure 1 about here

Gene Regulation

Dunbar (1993) devised another complex task environment, modelled on the discoveries of Monod and Jacob, in which subjects had to discover a mechanism for how genes are controlled. The mechanism involved the specialized functions of hypothetical inhibitor genes and mutant genes which could manipulate whether the inhibitor genes were present or absent. The subjects underwent a learning phase which predisposed them to thinking of genes as activating, or switching on, another gene, which was opposite to the actual inhibitory mechanisms that were to be discovered during the experimental phase. This bias towards seeking an activation role for genes proved difficult to shake, with thirteen of the twenty subjects still holding an activation mechanism hypothesis (with another three retaining some aspects thereof) at the end of the experiment.

One striking difference between those who deduced the inhibitory mechanism and those who retained the activation hypothesis was in how each used disconfirming evidence. Those subjects who did deduce inhibitory mechanisms used the disconfirming evidence to set up new goals which focused on trying to explain their discrepant observations. Those who failed to do so attempted to elaborate on the activation hypothesis to accommodate the disconfirming evidence, but in so doing were actually ignoring and/or distorting this evidence. Note that these discrepancies were first observed with tests designed to confirm the initial activation hypotheses. Dunbar speculated that the tester's goals were of central importance, since it is goals that determine where a line of enquiry goes, and consequently how successful the subject will be.

A second experiment yielded evidence suggestive of the early-late process mentioned above. In this experiment, it was postulated that providing subjects with some initial confirmation of the activation hypothesis would facilitate the change in the tester's goals. The argument runs that once the initial goals are satisfied (i.e., evidence is found in favour of an activation mechanism), one will focus on the discrepancies in one's data. Therefore, introducing an activation mechanism into the task should facilitate the discovery of the inhibitory mechanism, and this, in fact, is what was observed. In this experiment subjects had a working knowledge of the task at hand. They also received support for that initial hypothesis of an activation mechanism. With task knowledge and a working hypothesis, subjects can be said to be at a late stage of the hypothesis testing process. And it is under these circumstances that subjects will be more likely to attend to disconfirming or discrepant evidence.

Confirmation Bias as a Positive Test Strategy

The preceding sample of studies suggests that a strong predisposition to hypothesis confirming tests is common, but that this predisposition can sometimes be detrimental and sometimes not. Indeed, the implication that "confirmation bias" is maladaptive has been disputed by Klayman and Ha (1987), who proposed "that many phenomena of human hypothesis testing can be understood in terms of a general positive test strategy. According to this strategy, people test a hypothesis by examining instances in which the property or event is expected to occur (to see if it does occur), or by examining instances in which it is known to have occurred (to see if the hypothesized conditions prevail)" (p. 212). Klayman and Ha distinguished between a positive hypothesis test (+Htest) and a negative hypothesis test (-Htest); a +Htest is a test of a hypothesis with an observation that one expects to be a target (e.g., in the 2-4-6 task proposing the triple 8-10-12 would be a +Htest of the hypothesis, "increasing in twos"), whereas a -Htest is a test that one does not expect to be a target (e.g., 2-4-7).

Klayman and Ha (1987) made the case that under some very common conditions the probability of receiving falsification with +Htests can be greater than with -Htests. For example, the +Htest in the previous paragraph, 8-10-12, would falsify the hypothesis, "increasing in twos," if the true rule was "any three numbers less than 10." They also argued "that when concrete, task-specific information is lacking, or cognitive demands are high, people rely on the positive test strategy as a general default heuristic" (p. 212). As with all general heuristics, while there are particular situations in which the positive test strategy may lead to problems, "this general heuristic is often quite adequate" (p.212).

What are the implications for the pursuit of science? Scientific problems are typically such that concrete, task-specific information is lacking and cognitive demands are high. Klayman and Ha argue that "in probabilistic environments, it is not even necessarily the case that falsification provides more information than verification. What is best depends on the characteristics of the task at hand," (p.225), and that the "positive-test strategy" can, under realistic conditions, "be a very good heuristic for determining the truth or falsity of a hypothesis" (p.211).

Similar reanalyses of other reasoning tasks have stressed the optimal and adaptive nature of human reasoning (Oaksford & Chater, 1994) and have critiqued the unrepresentativeness of many laboratory tasks that have raised doubts about human rationality (Gigerenzer, 1996; Gigerenzer, Hoffrage & Kleinbolting, 1991).

The Confirmation Heuristic and the Early /Late Hypothesis

Kareev and Halberstadt (1993) addressed the issue of how people evaluate the potential benefits of negative tests and disconfirmations in the 2-4-6 task. They asked subjects to evaluate the usefulness of possible triples generated by imaginary subjects. In this way, the cognitive load in performing on the standard 2-4-6 task was lessened, which enabled the subjects to concentrate on the outcomes of each of the possible triples. They found that subjects evaluated positive tests of a hypothesis most highly. However, in a second experiment, in which subjects were aware of the true rule, a very different picture emerged. Giving subjects knowledge of the outcomes of the various tests (i.e., confirming or refuting) resulted in subjects judging refutations as being more useful than confirmations. Negative tests that refuted the hypothesis under test *and* revealed a positive instance of the true rule were deemed most useful.

Apparently subjects judge both refutations and positive instances of the rule as important. Refutations, which may result from either positive or negative tests, will identify the current hypothesis as being incorrect. Positive instances of the true rule, which may result from confirmations or disconfirmations, provide valuable information which may facilitate the generation of new hypotheses (including the correct one). Kareev and Halberstadt asked "Is it not possible that in the early stages of a scientific undertaking the principal goal is to amass a large stock of positive instances from which the learner can begin to distill patterns of contingencies?" (p.22).

The Present Study

The original conception of confirmation bias as maladaptive may not, for a variety of reasons, be appropriate in every situation. Under certain circumstances, especially in complex environments, the attempt to confirm one's theories, especially at an early stage in theory development and refinement, may be more efficient and effective. In the early stages of trying to discover or clarify some unexplained phenomenon, the task can be so difficult, and any information that comes to hand so confusing, that to proceed by trying to establish multiple hypotheses and test them via a falsificationist strategy may be beyond individual human cognitive capacities. Indeed, even trying to think of an assertion as false when one has been predisposed to thinking of it as true has been shown to be particularly difficult in a truthteller-liar task (Byrne, Handley & Johnson-Laird, 1995).

The present experiment will assess the efficacy of a multiple hypotheses and disconfirmation approach, at both early and late stages of trying to discover the rules of a more complex version of the Artificial Universe task (Mynatt et al., 1978). Our purpose is two-fold; to determine if a strict adherence to a falsification approach facilitates or impedes success on this rule-discovery task, and to determine if the impact of the falsification approach is affected by the level of one’s knowledge of the task.

METHODOLOGY

Subjects

Twenty undergraduates, all experienced with Macintosh computers, were randomly assigned by blocks to one of four conditions. There were 13 males and 7 females, ranging in age from 18 to 35 (mean age: 20.5 yrs). Half of the subjects were majoring in biological sciences, six in physical sciences and four in mathematics and computer science. They reported having had an average of 3.32 computer classes, and played video games, on the average, for 2.49 hours per week. One other subject was replaced as she had repeatedly reported being tired, hungry, and unmotivated during the experiment. Subjects received credit toward a research requirement and ten dollars for their participation.

Apparatus

The Artificial Universe is programmed in Microsoft Basic and runs on a Macintosh computer. The dimensions of the Universe are 2,555 screen units by 1,605 screen units, twenty-five times the area of a Macintosh Plus screen. One can view the entire Universe on a reduced single screen display. From this reduced display one can select any 510 X 320 portion of the Universe as a window in which to conduct experiments. The entire Universe contains 55 geometric objects, varying in shape, size and brightness, presented against an all-black background. The objects are of four shapes (squares, circles, rectangles and ellipses), two sizes (large and small), and of three brightnesses (white, black with white boundaries, or grey). All but the grey objects are surrounded by invisible circular boundaries with a radius of 150 screen units. These boundaries will henceforth be referred to as force fields. Though this term is, strictly speaking, incorrect, it will be used as subjects often use it when trying to explain the effects of these boundaries (see the Appendix for a complete description of the rules of the Universe).

Using a mouse-controlled cursor one can specify any locus in a window from which to "fire" a "particle," then enter an angle of fire. Then, using the mouse, one traces the path that one expects the fired particle to take. When this predicted path is completed, the particle originates its travel from the firing location in the direction specified and leaves a trace of its path. This entire sequence of operations is termed an experiment. An example of an experiment is given in Figure 2. Here, the particle was fired at a 0-angle (straight up) from the lower centre of the screen. The predicted path is drawn in white and the particle's actual path in grey, with the first few inches of the white trace covered by the subsequent grey one. As can be seen, the particle soon deviated from the predicted path turning first to the right and then to the left. Particles are deflected away from any object around which a force field is located, if and only if the particle path originates from outside the field. The force field surrounding the white square in the upper-left corner of the screen caused the first deflection while an object not on this screen caused the second deflection. This feature of the Universe task, that a force field can encroach on a screen though the object it surrounds does not, adds to its difficulty. A subject's task is to discover the rules governing particle motion by conducting experiments. The program requests the subject to type in his/her current hypothesis and to provide a confidence rating in this hypothesis after every ten experiments and records the details of all experiments.

Insert Figure 2 about here

Procedure

Subjects worked individually on the computer in the presence of the experimenter. Preliminary interaction with the experimenter was kept to a minimum as all instructions and manipulations were conveyed via computer. Each subject had his/her own computer diskette which contained all instructions, manipulations and the Artificial Universe task itself. The diskette recorded the typed responses to the different questions that they answered throughout the experiment as well as their actions on the Universe task. The subjects were invited to keep an informal account, on paper, of their thoughts, hunches, insights and so on as they worked on the task. Subjects were also encouraged to think-aloud and were tape-recorded.

Multiple Hypotheses/Disconfirmation Instructions

After the subjects read introductory instructions and supplied demographic details, they read an explanation of hypothesis testing and a description of what the experiment required of them. Then, all subjects were given a programmed version of the 2-4-6 task. Half the subjects (the INSTR group) then read instructions about both the advantages of generating multiple hypotheses and the advantages of a falsification strategy. These instructions contained the true rule (i.e., numbers must be in ascending order) which they may or may not have discovered. These subjects were then instructed to return to the 2-4-6 task where they were to test more triples using this disconfirmation approach. It was also explained to them that they should try to adopt this approach when working on the Universe task.

To practice their understanding and comprehension of these instructions the INSTR group completed another triples-type problem based on one used by Gorman and Gorman (1984). This task included three triples, a hypothesized rule and instructions to test the truth of the rule with more triples. The INSTR subjects were asked both to provide three new alternative hypotheses and to test these hypotheses with triples using the disconfirming strategy.

A similar set of instructions given to the control subjects (CTRL) contained no mention of multiple hypotheses or disconfirmation. Given prior research that suggests that a confirmation or positive test strategy is a default heuristic, we anticipated that these subjects would adopt such a confirmatory strategy, though they were not explicitly instructed to do so. To control for time spent on instructions across conditions, the subjects in the CTRL conditions worked on a similar task which asked for three triples to test a particular rule, no mention being made of generating alternative hypotheses or of trying to falsify hypotheses.

All subjects read instructions on how to use the Artificial Universe program, that is how to change screens, carry out experiments, record their current hypotheses and so on. They were then given ten minutes to familiarize themselves with the task. Subjects in the CTRL conditions next received instructions on how to test hypotheses on the Universe task. The subjects in the INSTR conditions were given an example of how to use the disconfirming strategy in the Artificial Universe. Their version of instructions on how to test hypotheses in the Universe stressed the
generation of alternative hypotheses and the value of disconfirmation.

The instructions for both the INSTR and CTRL conditions included a practice hypothesis, namely "that all the particles are being attracted to the very centre of the Universe." CTRL subjects were told to test this in order to practice the procedures for testing hypotheses in the Universe. INSTR subjects were directed to put into practice the instructions they had received by trying to disconfirm this hypothesis. These latter instructions explained that one could disprove this hypothesis by firing away from the centre of the Universe from many different places on many different screens in the Universe. If the particles did not turn back in towards the centre of the Universe then one could interpret this as evidence that disconfirmed the hypothesis. If they did turn back in towards the centre then one would have confirmatory evidence. The instructions continued that one could also disprove this hypothesis by firing towards the centre of the Universe in the hope of seeing the particle being deflected away from the centre. If the particles did deflect away from the centre then one would have disconfirmatory evidence. These instructions also suggested to subjects that before they attempted to test this practice hypothesis they should spend time in first trying to think of other possible hypotheses that might better explain what is happening to the particles in the Universe.

All subjects were then instructed to spend a number of minutes working on the Universe task with the hypothesis "that all the particles are being attracted to the very centre of the Universe." The entire set of instructions to subjects were quite lengthy. They may be viewed by readers on the Web page <http://www.bgsu.edu/????>.

The Early/Late Stage Manipulation

A late stage manipulation for half of the subjects in the INSTR condition and half in the CTRL condition was created by providing a large-scale map of the entire Universe. This map, measuring 87.6 cm X 57.2 cm, was on a display board to the side of the desk on which the computer containing the Universe task sat. It was formed by taking 25 screen dumps from the Universe, and juxtaposing adjacent screen dumps so as to represent faithfully the entire Universe. Each screen displayed the results of an experiment, that is a firing angle and the subsequent particle path. The firing locations and firing angles were randomly chosen.

The distinction between the late stage and the early stage of the hypothesis testing process is defined by the presence or absence of this map of prior experiments; it was available throughout the experiment for subjects in the MAP condition but was never shown to subjects in the NO-MAP condition. These prior experiments were intended to provide a knowledge base concerning the Universe at the commencement of the task. Showing the Universe in its entirety was intended to yield insights into its integral nature (e.g., that objects not on the screen in view could still influence particle motion on that screen; see appendix rule 6), suggest hypotheses, and provide data relevant to those hypotheses. Completing all instructions, to the point just before the subjects started on the Universe task itself, took approximately one hour.

The Universe Task

All subjects worked on the Universe task for two hours. Throughout that time subjects occasionally (perhaps once or twice) received prompts to formulate hypotheses were they neglecting to do so. Subjects were told that formulating a hypothesis, indeed any hypothesis, would aid them in trying to come to terms with the task. When called upon to type in their current hypothesis, subjects in the INSTR conditions were also prompted to generate alternative hypotheses. On providing their current hypothesis all subjects were asked to elucidate their immediate plans to either test or disprove this hypothesis depending on the condition that they were in. These protocols enabled the experimenter to categorize, however imperfectly, the type of testing that was to follow, for example, as +H or -Htests. Subjects were allowed to take a break at any stage throughout the two hours. Time taken on breaks was not included in the two hours on the Universe task.

After two hours working on the task and after having received a notice that five minutes remained, the subjects were asked to stop conducting experiments. Subjects in the INSTR conditions then completed a short self-report measure on the computer with respect to how closely they believed that they had adhered to the approach that they were instructed to adopt, that is whether or not they took time to generate alternative hypotheses and whether or not they tested these using the falsification strategy. At this stage all subjects had completed their time working on the computer and the Universe program was exited.

Measures of Success

There were three measures, completed in order of increasing reactivity. They were: (a) An open-ended question in paper-and pencil format in which subjects wrote down what they considered to be the rules that govern particle motion. They were allowed to refer to their written notes when answering this question. These notes were then collected, and the map that had been present for subjects in the MAP condition was turned out of view. (b) Fifteen scenes, presented on individual sheets of paper, similar to those that one would find within the Universe. For example, a typical scene might contain a black square and a white rectangle. Each scene showed the results of a prior experiment, i.e., the firing location, firing angle and particle path were all shown. The subjects' task was to decide whether the particle path shown was the correct one for that firing location and firing angle. If the subject believed it to be correct they were to write TRUE on the paper; if they thought that it was incorrect then they were to write FALSE and draw what they believed should be the correct path. Each scene was constructed to test the subjects on a particular rule of the Universe. (c) A multiple choice test administered in two parts: part one sought to discover the subjects' understanding of the rules of the Universe while part two presented the subjects with a correct, though incomplete, description of the Universe and asked more specific questions related to the Universe's more detailed rules. These tests are also available at <http://www.bgsu.edu/????>.

On completing this final test the experimental session was concluded and subjects were debriefed on the rules of the Universe. In total, the experimental session lasted about four hours. The first measure, the written explanation of the rules of the Universe, was scored by two independent raters who were blind to the experimental condition. Each of the rules of the Universe had been assigned a certain value (see Appendix) and the written explanations were awarded these values for each rule correctly identified. These two raters also assessed the accuracy of the subjects on the scenes from the Universe. Drawings which were correctly identified as FALSE were scored correct only if the path that the subject then drew was deemed sufficiently accurate. The measure taken from the multiple choice test was simply a count of the number of correct answers.

Design

A full factorial, between-subjects design was employed. Half of the subjects received the instructions on multiple-hypotheses testing and disconfirmation (INSTR), the other half received the control instructions (CTRL). This manipulation was crossed with the presence of the map, half of the subjects had the map present throughout the experiment (MAP), the other half never saw the map (NO-MAP). Thus, the 2 (Instructions) x 2 (Map) design created four cells with 5 subjects per cell. It was decided that concern for low statistical power with so few subjects could be justified by the potential for qualitative insights that testing fewer subjects individually would provide.

RESULTS

Quantitative results

The 20 subjects carried out from 39 to 197 experiments with a mean of 120 and a standard deviation of 36. Inter-rater reliability for the sum of the scores for the three measures was r = .96. Scores on each measure for each rater were converted to z-scores, and then averaged across the two raters. These averages were in turn converted to z-scores, and will be referred to as test scores. These conversions allow for equal weight to be given to the two raters and for equal weight to be given to the three tests. The sum of these three test scores will be referred to as the total test score.

A 2 (Instructions) x 2 (Map) ANOVA was performed on the total test scores. The subjects in the Control conditions were more successful on the task than subjects in the Instructions conditions, F(1, 16) = 4.88, p = .04. Neither the Map manipulation nor the interaction proved significant (F's < 1). Additional 2 x 2 ANOVAs performed on the individual test scores yielded similar results. Only the Instructions effect yielded F-values greater than 1 on each test (Essay: F = 3.93, p = .07; Screens: F = 5.76, p = .03; MC test: F = 2.55, p = .13) with the CTRL subjects scoring higher than the INSTR subjects in each case. Raw means for each of these tests are provided in Table 1. Analyses performed on these raw scores produced an identical pattern of results. An additional looser criterion was applied to the Screens task. Particle paths that were correctly identified as FALSE were scored correct irrespective of the accuracy of the corrected path drawn by the subject. ANOVA on this looser criterion once again showed superior performance by CTRL subjects, F(1, 16) = 5.75, p = .03.

Insert Table 1 about here

Table 2 shows the number of rules discovered by the subjects in each condition. The correlation between the number of experiments that the subjects performed and their total test score was -.44 (p = .05). The Discussion section is predicated in part on the analyses of the test scores and in part on the qualitative results that are described below.

Insert Table 2 about here

Qualitative results

Requiring subjects to work on the task individually and to think aloud in the presence of the experimenter yielded a rich body of qualitative information. This information combined with the subjects' lab notebooks allowed for an idiographic report of each subject's performance on the task. These reports are available from the authors. In the following paragraphs we try to convey some sense of the quality of the participants' behaviour.

Across subjects, numerous hypotheses were generated. The firing angle and firing location were proposed as sole determinants of the particle paths. Invisible lines existing between objects, gravitation and repulsion effects, and mirrors in which the paths in one portion of the Universe were reflected in other portions were also hypothesized to exist. Skeptical hypotheses were advanced, with one participant expressing the possibility that there existed a predetermined sequence of paths, a sequence that was independent of any observable antecedents. Extremely complicated hypotheses were also entertained, with particle motion being, for example, a function of interactions between grid coordinates and firing angles.

Subjects employed a variety of strategies in testing these hypotheses. It was common for subjects to try to hold all aspects of the task constant while varying just one feature, for example firing from the same location but with different firing angles, or firing at the same type of object from a constant angle but on different screens. This is essentially the standard logic of experimentation, at least as taught in introductory classes, and is a quite reasonable strategy for these subjects. Another common strategy was to follow a particle across screens, firing within an adjacent screen from the particle's point of entry and at the particles approximate angle. Others, thinking that it might reveal something interesting, only fired at objects that were off-screen, or only fired at objects from behind other objects. It was also common for subjects to make drawings of their experiments on paper, and one subject insisted on ten replications of every experiment he performed. Interestingly, confusion was created by sloppy experimentation; some participants, attempting to replicate an experiment, and incorrectly thinking that they were at the same firing location, found it difficult to explain the seemingly inconsistent results that were being observed.

Other aspects of scientific inference were observed. Use of analogy was evidenced, with "force field," "planet" and "gravitation" metaphors being common, and objects being hypothesized to emit waves or energy. There were occasions in which an analogy created difficulties. One subject, believing that the visible objects were in fact the centres of larger, same-shape objects, attempted to explain the deflections of the particles in terms of a ball bouncing off a solid physical surface. However, in investigating a rectangular object, he was at a loss to explain what were in fact tangential deflections (see Appendix, rule 5) off what he thought was a straight boundary. Evoking the ball metaphor, he hypothesized that the particle might be spinning; hence the unexpected deflections. Trying to verify this consumed time. Indeed, when he was unable to find the necessary verification, he then hypothesized that perhaps it was the objects themselves that were spinning.

Also observed was a phenomenon that has often been proposed as playing a role in scientific inference, that is, serendipity. One subject was observing what she considered inconsistent results from what she thought were identical experiments. Careless experimentation, however, had led her to stray in and out of the force field of an object, and consequently she was observing deflections only some of the time. This caused her much anguish, as she felt that she had a good understanding of the Universe. Unfortunately her understanding did not include the importance of distance from the objects, which became clear to her only when she accidentally set her firing location much further away from the object than she had intended. In noting the resulting path, the connection between distance from the object and deflection became salient.

DISCUSSION

The Effects of Using Multiple Hypotheses and Falsification

Subjects who received instructions to generate and attempt to disprove multiple hypotheses appeared to follow them when working on the Universe task. When reporting both how often they had attempted to generate alternative hypotheses and how often they had employed the falsification approach, modal responses were sixes on the self-report seven-point scales (1 indicated “almost never” and 7 indicated “almost always”). It also appeared, from the 2-4-6 manipulation check, that these subjects had understood clearly what was meant by generating alternative hypotheses and seeking disconfirmations. Thus, subjects seemed to adhere to the Popperian ideal. However, doing so degraded performance!

Why was performance worse with such a strategy? One possibility is cognitive overload; perhaps concentrating on tests that could disconfirm a hypothesis is simply a more difficult task than concentrating on testing a hypothesis. Whereas subjects in the CTRL condition were free to test hypotheses in whatever way they chose, subjects in the INSTR condition were trying to devise tests that were specifically disconfirmatory. As well as being a more difficult requirement for the INSTR subjects this might also have limited their freedom to tackle the task in whatever way they saw fit.

Another consequence, perhaps arising from a need to reduce the cognitive load, is that the effort to disconfirm may have focused a subject’s attention too narrowly. It appeared that some of the INSTR subjects construed their task differently from subjects who had received no such
instructions. They seemed to concentrate solely on formulating and disconfirming hypotheses as if that were an end unto itself, as opposed to being the means to another, more fundamental end, namely understanding the Artificial Universe. These subjects consequently became intent on their pursuit, failing, for example, to take time to reflect on the task and to allow their imaginations to suggest possible explanations or analogies. They appeared to forget their primary objective, to discover the rules of the Universe. This conjecture is in keeping with Dunbar's (1993) assertion that the goals subjects set for themselves can determine their behaviour on a task.

Finally, INSTR subjects often appeared to overweight disconfirming evidence. It was not uncommon for them to have their progress impeded by prematurely concluding, in the light of disconfirming evidence, that their current hypothesis was incorrect. For example, S#3 at one stage suggested that there existed something surrounding a white rectangle. In seeking to disconfirm this hypothesis he fired through a white rectangle from within its force field. After firing once more through the rectangle and then through a grey object, S#3 dismissed this (potentially fruitful) line of thinking.

Thus three factors, added cognitive demands, narrowing of focus and goals, and overweighting of disconfirming evidence may have contributed to the poorer performance of the INSTR subjects. This analysis is based largely upon the experimenter's observations of the subjects as they performed the task, not systematically collected data, but it should be amenable to more rigorous investigation.

The Lack of an Early/Late Effect

It appears that the large scale map of the Universe failed to create a "late" stage in the hypothesis testing process. The overt behaviour of many subjects in the MAP conditions may help explain this. We predicted that the experiments displayed on the map would facilitate the generation and disconfirmation of hypotheses, as well as providing subjects with insights into the integral nature of the Universe. However, it appears that the map was not sufficiently salient to the subjects. Though some did study it before embarking on their own experiments, they tended to ignore it once they had become absorbed in the task. It would seem that they focused entirely on the dynamic information that was coming to them from their own experiments. In fact, it was not uncommon for subjects to draw the results of their experiments in their journals and study these, all the while ignoring the twenty-five experiments available on the board. Indeed, some hypotheses could have been disconfirmed with a glance at the map. Two and perhaps three of the ten late-stage subjects ignored the map completely. One wonders if this may be related to the often informally reported tendency of scientists not to believe a result until they have replicated it themselves.

Success on the task

A prerequisite to solving this task is the development of a causal model that contains appropriate elements. To explain, there are many features to the Universe, all of which might appear to be likely candidates for the cause of particle motion. As noted above, the subjects as a group entertained hypotheses about a rich variety of possibilities. The key to success seemed to involve having the correct variables in one's model *before* becoming wedded to the wrong variables.

Take, for example, a subject who decides early on that the objects in the Universe appear to have some sort of effect on the particles. Though such a subject is unsure of just what the effect is, or how it works, or whether all the objects affect the particles, he or she nonetheless has a potentially appropriate causal model of the Universe, one that can productively accommodate subsequent observations. This issue of determining the correct variables has been addressed by Klayman (1988) with respect to cue-learning, specifically to learning from outcomes. Klayman asserted that there are several different things that one must do in order to learn, and that the first of these is determining what the important variables in a situation are. This he added "is perhaps the most poorly understood aspect of learning from experience" (p.116). He reported that providing outcome feedback does make it possible to discover the importance of certain cues over others, and that subjects do better at this when designing their own experiments. Note that the Artificial Universe task incorporates both of these features into its design.

Arriving at an appropriate model, however, is not a straightforward task. Take, for example, one subject who observed a deflection caused by a white rectangle which sent a particle to the upper-right hand corner of the screen on which he was working. This subject then concluded that the location to which the particle was deflected was all-important and it was this irrelevant factor on which he then concentrated. Had this subject carried out more experiments on this white rectangle he would most likely have realized that his conclusion did not generalize and that the causal element was in fact the white rectangle. But of course, for this subject to have performed more experiments on this object, the object itself would have to have been already prominent in his causal model. Perhaps an interesting way to characterize problem-solving behaviour and problem-solving progress is through the movement of variables in and out of the causal model as well as up and down, in terms of prominence, within it. Note that the usage here implies that a given causal model can support many different hypotheses; objects can attract or repel, but at least the subject has the appropriate element, that is, objects, in his or her model.

Though working on the task with a causal model which at least has the right elements in it is no guarantee of success, it is evident that a subject who has such a model is more likely to be successful than is a subject who believes, for example, that it is the firing angle that determines particle directions. One who has such a firing angle model in mind would not attend to the relations between deflections and objects, nor think to ask about the difference between grey objects and white and black ones. One who believes that the firing angle is crucial can only be successful when this model is abandoned.

Changing causal models

A particular causal model was typically abandoned only after much disconfirming evidence, or in the light of evidence which could not be accommodated by the model. Following abandonment of a model, subjects often engaged in a period of apparently hypothesis-free data gathering in which they attempted to reacquaint themselves with the task without the blinkers of the former model. In other words, at times the subjects appeared to be making observations in the service of generating hypotheses, rather than testing, confirming, or falsifying hypotheses. These periods were characterized by evident frustration on the part of the subject, followed by the adoption of a new causal model. The adoption of a new model appears to be similar to acquiring an insight, in that it usually appeared suddenly, and full-blown.

The pivotal role of an appropriate causal model was made all the more apparent in the present investigation by virtue of the fact that subjects were free to arrive at any model of their choosing, albeit perhaps by an unconscious choice. Unlike, for example, Dunbar's (1993) gene-regulation task, in which subjects were predisposed by instructions to a certain model, our subjects could choose their own; some chose well, others did not. A poor choice, made perhaps in the first minute or so, often proved an enormous hindrance to subsequent success on the task, whereas a correct choice, perhaps a fortuitous one, rendered the task far easier to solve.

Research on mental models may provide a potentially fruitful approach to understanding what determines a subject's choice. di Sessa (1983) proposed that physics-naive students see and explain the world with a repertoire of recognizable phenomena, which she labeled phenomenological primitives. Thus, the bouncing of a ball may be explained in terms of its inherent "springiness," a quality that is taken as primitive in the sense of requiring no further explanation. These primitives are "recognized" as being relevant for a particular task or situation and it is through these primitives that the student arrives at an understanding of the phenomenon of interest. Based on the person's experience and knowledge these primitives have preexisting priorities that determine the probability and duration of their usage. Thus, a useful question becomes, with what primitives do subjects perceive the Artificial Universe task? Clearly, the way the task is presented, the subject's background experience and knowledge, and indeed any biasing instructions (e.g., to employ a certain analogy) could influence a subject's perception, and maybe success, on the task. Some phenomenological primitives that all subjects used to explain the task, and indeed that governed the design of the task, were the object quality of the features and the particles, and the causal interactions among them (reminiscent of the early work of Michotte, 1963), as well as the idea of forces, both attractive and repulsive. Other, more idiosyncratic primitives, were undoubtedly at the root of some of the individual differences between subjects.

Types of Hypothesis Tests: An Elaboration of the Early-Late Proposition

This part of the Discussion is based in part on the direct observations made on the subjects by the experimenter, as they "thought aloud" during the task, and on the subjects' notes. Subjects sometimes seemed to be gathering information with no clear hypothesis-related purpose at all. We think that this was because they had no hypothesis, or at least none that had implications for specific tests. This type of behaviour typically happened early in the experiment, and in the interval between the abandonment of one hypothesis and the development of another. Let us call such tests Information tests, or Itests, in which people are (apparently) looking for information with which they can formulate a hypothesis.

Moreover, even when subjects did have some sort of hypothesis, many of their tests were neither +Htests nor -Htests. Rather, they fell in a grey area between pure information gathering with no hypothesis in mind and a distinct test of a specific hypothesis. That is, many tests seemed to be performed for the purpose of fleshing out a vague, inchoate hypothesis, rather than testing one. Take for example, the hypothesis that the objects seem to have some role in particle motion. One might test this by gathering information to see if there exist patterns of particle motion in the vicinities of the objects. One might wish to see if these patterns are the same for all objects or only for certain subsets of objects. These tests are not truly hypothesis-free, for the subject does have expectations about how the results should turn out. However, the subject is not testing a specific hypothesis, for at this stage his or her knowledge and thinking is not yet sufficiently complete to allow a strong test of a hypothesis.

A test of this sort cannot be accurately described as either a +Htest or a -Htest, for how can one test for an effect where one expects it to occur, if one is unsure of just what the effect is? We propose that one might usefully characterize performance on a hypothesis testing task in terms of a progression, one marked by stops, starts and even setbacks, in the types of tests that people perform and that these types of tests are representative of different stages in hypothesis testing. Imagine how subjects might behave on encountering a difficult rule discovery task in which they have no previous experience. They would have to embark on some information-gathering in order to become familiar with and arrive at some sort of rudimentary understanding of the task. Though they will certainly bring their own background knowledge and experience to the task, we would still consider this an information-gathering stage. They may then go through a stage of trying to clarify their thinking on the task, a stage which would be characterized by a more purposeful information gathering, directed by an indistinct or inchoate hypothesis. Finally, on arriving at a formal hypothesis they might attempt to test it, presumably with a preponderance of +Htests, with the intent of yielding either confirmations or disconfirmations.

Each of these stages would be demarcated by a predominance of certain types of tests. We suggest that these types may be categorized as:

Itests Information-gathering with the goal of formulating a hypothesis

+HItests Information-gathering with a hypothesis-related intent, but where the hypothesis is vague or generic

+Htests Testing for an effect where one is predicted to occur, or testing cases that are expected or known to have the property of interest (Klayman & Ha, 1987)

-Htests Testing for an effect where one is predicted not to occur, or testing cases thatare expected or known to lack the property of interest (Klayman & Ha, 1987)

The first three types of tests may be thought of as existing on a continuum. The boundaries between them are indistinct insofar as labeling a particular test as pure information-gathering or information-gathering with some vague hypothesis may often be very difficult. Where a particular test falls along the continuum should be determined by the extent of the tester's knowledge of the task. The fourth type of test, -Htests, would seem to be a more distinct type of test. Though it too is often difficult to identify and would seem to be related to the knowledge of the tester, it would not seem to overlap with the other types of tests. Though this speculation has been couched in purely descriptive terms, there are prescriptive implications, much as there were with the original early-late proposition.

Implications

This study has demonstrated that an adherence to a falsificationist approach to hypothesis testing proved deleterious to performance. While falsifiability may be a useful demarcation rule, this research suggests that the attempt to falsify is not a good prescription for science. This has been demonstrated in an experimental task with a content sufficiently complex for subjects to generate a rich array of metaphors, analogies, and hidden mechanisms in trying to understand it. Further, to better describe the variety of tests that are performed on such a task, the intentions of the tester must be understood. The category system of hypothesis tests proposed by Klayman & Ha (1987) appears to be very useful for understanding hypothesis testing in well-defined tasks, but it has to be expanded to describe hypothesis testing in more open-ended domains.

The most interesting outcome of the present research is the deleterious effect of the falsification instructions. We emphasize that this effect occurred in an environment that was, in theory and practice, perfectly predictable. We surmise that this effect would be exarcebated by the presence of the sort of measurement error and environmental uncertainty that characterizes both the scientist's struggle to understand results in the laboratory and everyone's struggle to understand causality in the world.

ACKNOWLEDGEMENTS

This manuscript was prepared with the support of National Science Foundation grant SBR-9422253 to Bowling Green State University, Michael E. Doherty and Clifford R. Mynatt principal investigators. The authors wish to gratefully acknowledge comments from Ruth Byrne and Roddy Cowie.

REFERENCES

Byrne, R. M. J., Handley, S. J. & Johnson-Laird, P. N. (1995). Reasoning from suppositions. *Quarterly Journal of Experimental Psychology*, 48(A) 4, 915-944.

Chalmers, A. F. (1982). *What Is This Thing Called Science?* University of Queensland Press: St. Lucia, Qld.

di Sessa, A. A. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. L. Stevens (Eds.), *Mental Models*. (pp. 15-33). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Doherty, M. E., Mynatt C. R., Tweney, R. D. & Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica*, **43**, 111-121.

Doherty, M. E., Chadwick, R., Garavan, H., Barr, D. & Mynatt C. R. (1996). On people's understanding of the diagnostic implications of probabilistic data. *Memory and Cognition,* **24**, 644-655.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, **17**, 397-434.

Eccles, J. C. (1981). In praise of falsification. In R. D. Tweney, M. E. Doherty, & C. R. Mynatt (Eds.), *On Scientific Thinking* (pp. 109-110). New York: Columbia University Press.

Einhorn, H. J. & Hogarth, R.M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, **85**, 396-416.

Evans, J. St. B. T. (1989). *Bias in Human Reasoning*. London: Erlbaum.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. Psychological Review, 103 (3), 592-596.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, **98 (4)**, 506-528.

Gilovich, T. (1991). *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*. New York: Macmillan Inc.

Gorman, M. E. & Gorman, M. E. (1984). A comparison of disconfirmatory, confirmatory and a control strategy on Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology*, **36A**, 629-648.

Gorman, M. E., Stafford, A. & Gorman, M. E. (1987). Disconfirmation and dual-hypotheses on a more difficult version of Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology*, **39A**, 1-28.

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.

Kareev, Y. & Halberstadt, N. (1993). Evaluating negative tests and disconfirmations in a rule discovery task. *Quarterly Journal of Experimental Psychology*, **46A**, 715-727.

Kern, L. (1982). The effect of data error in inducing confirmatory inference strategies in scientific hypothesis testing. Unpublished Ph.D dissertation, The Ohio State University.

Kern, L. H., Mirels, H. L. & Hinshaw, V. G. (1983). Scientists' understanding of propositional logic: An experimental investigation. *Social Studies of Science*, **13**, 131-146.

Klayman, J. (1988). On the how and why (not) of learning from outcomes. In B. Brehmer & C.R.B. Joyce (Eds.), *Human Judgment: The SJT View*. Holland: Elsevier Science Publishers B.V.

Klayman, J. & Ha, Y.(1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, **94**, 211-228.

Kuhn, T.S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lakatos, I. (1970). Falsification and methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Scientific Knowledge,* (pp.91- 196). New York: Cambridge University Press.

Lakatos, I. (1978). History of science and its rational reconstructions. In J. Warrall & G. Currie (Eds.), *The Methodology of Scientific Research Programmes.* Philosophical papers of Imre Lakatos (Vol. 1, pp. 102-138). Cambridge, England: Cambridge University Press.

Mahoney, M. J. & DeMonbreun, B. G. (1977). Psychology of the scientist: An analysis of problem solving bias. *Cognitive Therapy and Research*, **6**, 229-238.

Michotte, A. (1963). *The Perception of Causality.* New York: Basic Books.

Mitroff, I. (1974). *The Subjective Side of Science*. Amsterdam: Elsevier.

Mynatt, C.R., Doherty, M.E. & Tweney, R.D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology*, **30**, 395-406.

Oaksford, M. & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, **101 (4)**, 608-631.

Platt, J.R. (1964). Strong inference. *Science*, **146**, 347-353.

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

Quine, W. V. O. (1961). Two dogmas of empiricism. In W. V. O. Quine (Ed.), *From a Logical Point of View*, (2nd ed., pp. 20-46). New York: Harper & Row. (Originally published, 1953).

Sekuler, R. (1976). Seeing and the nick in time. In M.H. Siegel & H.P. Zeigler, (Eds.). *Psychological Research: The Inside Story.* New York: Harper & Row.

Tweney, R. D. (1989). A framework for the cognitive psychology of science. In B. Gholson, W. R. Shadish, Jr., R. A. Neimayer & A. C. Houts (Eds.), *Psychology of Science: Contributions to Metascience*, (pp.342-366). Cambridge: Cambridge University Press.

Tweney, R. D., Doherty, M. E. & Mynatt, C. R. (1981). *On Scientific Thinking*. New York: Columbia University Press.

Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A. & Arkkelin, D. L. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology*, **32**, 109-123.

Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, **12**, 129-140.

Wason, P.C. (1968). On the failure to eliminate hypotheses-A second look. In P. C. Wason & P.N. Johnson-Laird (Eds.), Thinking and Reasoning, (pp.165-174). Harmondsworth, Middlesex, England: Penguin.

Wason, P. C. & Johnson-Laird, P. N. (1972). Psychology of Reasoning:Structure and Content, Cambridge: Harvard University Press.

APPENDIX

Rules of the Artificial Universe Task:

1. Deflections are caused by "force fields" or "shells" that surround most of the objects in the Universe. (8)[1]
2. Grey objects have no force fields. (5)
3. Force fields are circular and are centred on the centre of the objects. (3)
4. A force field is ineffective if a particle originates inside of it. (3)
5. The deflection of a particle will always be tangential to a force field. (2)
6. Force fields that reach into the active window, though the object with the force field does not, can still influence the paths of the particles in the active window. (2)
7. Force fields are operative for one and only one deflection per experiment, such that if a particle rebounds back towards a force field from which it has already deflected that force field will have no effect on the particle. (1)
8. Force fields have radii of 150 screen units. (1)

[1] The values in parentheses are the points awarded for correctly identifying the rule.

Table 1

	CTRL	INSTR
Essay	9.25 (2.57)	3.35 (1.02)
Screens (strict criterion)	9.85 (1.11)	6.50 (0.82)
Screens (loose criterion)	11.30 (0.97)	8.20 (0.81)
Multiple-choice test	15.85 (1.94)	12.00 (1.31)

Table 2

	CONTROL		INSTRUCTIONS	
Rule	Map	No Map	Map	No Map
Particle deflections are caused by FFs[1]	3	3	3	1
Grey objects do not have FFs	2	4	0	2
FFs are circular, centered on the object	2	1	0	0
FFs have no effect on particles fired from within FFs	3	1	0	0
Deflections are tangential to the FF	2	2	0	0
FFs outside the active window project into it	0	0	0	0
FFs deflect only once per experiment	0	0	0	0
FFs have a radius of 150 pixels	2	2	2	0
TOTAL	14	13	5	3

[1] The abbreviation FF means force field, as used in the paper. See appendix for more complete statements of the rules.

Table Captions

Table 1. Raw means and standard errors on each success measure.

Table 2. Number of subjects solving each rule, by condition.

Figure Captions

Figure 1. Nine-screen universe from Mynatt et al., 1978. Objects were of three brightnesses, shown as black, grey and white. Boundaries, which were not visible to subjects, are shown surrounding those objects that caused deflections.

Figure 2. Example of an experiment. Here, the particle's path, shown in grey, changed direction twice. A subject's task is to discover the rules that determine these paths.