Spring 1981

HOW COMMUNICATION AND CONFIRMATORY STRATEGIES AFFECT THE SEARCH FOR TRUTH

MICHAEL ERNEST GORMAN

8129267

GORMAN, Michael Ernest

HOW COMMUNICATION AND CONFIRMATORY STRATEGIES AFFECT THE SEARCH FOR TRUTH.

University of New Hampshire Ph.D. 1981

University
Microfilms
International 300 N. Zeeb Road, Ann Arbor, MI 48106

HOW COMMUNICATION AND CONFIRMATORY STRATEGIES AFFECT THE SEARCH FOR TRUTH

by

Michael E. Gorman
B.A., Occidental College, 1974
M.A., University of New Hampshire, 1978

DISSERTATION

Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy
in
Psychology

May, 1981

This dissertation has been examined and approved.

Daniel C. Williams

Dissertation director, Daniel C. Williams
Associate Professor of Psychology

David E. Leary

David E. Leary, Assistant Professor
of Psychology

R Michael Latta

R. Michael Latta, Assistant Professor
of Psychology

Donald M. Murray

Donald M. Murray, Professor of English

R Valentine Dusek

R. Valentine Dusek, Assistant Professor
of Philosophy

Date April 21, 1981

Dedication

To the memory of Mrs. Helen Post Hartz and the Hawkeye Trail Camps.

Acknowledgements

I would like to thank the members of my Ph.D. committee, especially David E. Leary and Donald M. Murray, both of whom played crucial roles in launching my academic career.

Thanks also to all my friends, particularly Michael Berry, David Bozak, Douglas Lea and Margaret Erwin, all of whom provided valuable assistance.

Finally, thanks to Arwen and Strider, without whose help this thesis would have been finished months ago.

Briefer versions of this thesis were presented to the Eastern Psychological Association in April of 1981 and the American Psychological Association in August of 1981. The work was funded by a University of New Hampshire Dissertation Fellowship. I am grateful to all these organizations for their support.

TABLE OF CONTENTS

ABSTRACT

HOW COMMUNICATION AND CONFIRMATORY STRATEGIES AFFECT THE SEARCH FOR TRUTH

by

Michael E. Gorman

University of New Hampshire, May, 1981

Scientific reasoning has become a topic of recent psychological research. Studies have focused on Karl Popper's idea that scientists should try to falsify, or disconfirm, their hypotheses instead of verifying them. Results indicate that both scientists and college students prefer to use confirmatory logic on simple tasks that model scientific reasoning. The only attempt to instruct subjects to use disconfirmatory reasoning failed (Mynatt, Doherty & Tweney, 1977) though subjects that falsified on their own initiative were more successful than subjects who tried to confirm.

All the studies of scientific reasoning have focused on individuals. But major advances in science are often made by groups, e.g., the research teams that discovered the structure of DNA and developed the atomic bomb. Experimental studies of group problem-solving have compared the performance of interacting groups with that of concocted groups composed of an equal number of individuals working separately. When there is a single, right answer to a problem, interacting groups perform about as well as the *best* member of each equally large concocted group, but better than the average person working alone.

This thesis synthesizes the literatures on scientific reasoning and group problem-solving by combining their two major variables in a single study. *Communication* was manipulated by running subjects in groups of four and either telling them to interact or to work separately. *Strategy* was manipulated by instructing subjects to follow either disconfirmatory or confirmatory approaches to the task, which was based on New Eleusis, a card-game designed to model the "search for truth." Each group had to solve the same four increasingly difficult Eleusis problems.

The overall design was a 2 (interacting vs. non-interacting) X 2 (disconfirmatory vs. confirmatory) X 4 (the Eleusis problems) split plot. Analyses-of-variance were conducted on the number of correct solutions and the time-to-solution achieved by groups in each condition.

Even though a manipulation-check revealed that disconfirmatory groups did try to follow their suggested strategy, there were no significant differences in the performances of confirmatory and disconfirmatory groups. This result replicates Mynatt et al.'s (1977) earlier research.

Interacting groups performed no better than the *best* member of each non-interacting group, where the best is defined as the person who solved each rule in the least time. Interacting groups also took significantly more time. But interacting groups did solve a significantly higher percentage of problems (80%) than all non-interacting individuals combined (33%). These results replicate earlier research on group problem-solving (Steiner, 1972).

A follow-up study, using the same task and interacting groups, revealed that disconfirmatory instructions produce superior performance when subjects have maximum freedom to design their own experiments. When the range of possible experiments is limited, confirmatory groups may serendipitously disconfirm their hypotheses.

A discussion of the implications of these results for science and suggestions for future research were included in the thesis.

INTRODUCTION

The focus of the thesis presented here is an experimental investigation of how two factors affect groups' attempts to solve a series of problems that model scientific reasoning. The paper is divided into four sections. The first section, "Studies of Scientific Reasoning," contains the background and rationale for the present study. The second section, "Methods," describes the details of the design and the third section, "Results," contains the quantitative findings. The final, "Discussion" section ties the results back in with themes presented in the first section.

I. STUDIES OF SCIENTIFIC REASONING

In the last 100 years, with the development of satellites, atomic energy, skyscrapers, radio and television, etc., the human environment has changed more than in the preceding 10,000. These changes have not affected all parts of the world equally: the developing nations have been less affected than the industrial ones. But most people now live in a very different world than the one their grandparents were born into.

Science has played a major role in all these changes. The classic example is Einstein's $E=MC^2$, a purely theoretical equation. It never occurred to Einstein that his formula could actually be applied to a technological problem. But at Hiroshima, Einstein's equation was translated into action--with horrifying results.

Science is not solely responsible for major advances in technology. Animal husbandrists had developed excellent breeding techniques long before Darwin and Mendel explained how their techniques worked. But the more recent discovery of the structure of DNA has opened the way to a whole set of applications that would have been impossible without that major scientific advance. Sometimes the scientific advance follows the technological one--but usually the order is reversed.

To understand the new world created by technology, it is necessary to understand science. Modern scholars in various disciplines have begun to study science in earnest. Gerald Holton (1973), Thomas Kuhn (1970) and others have thoroughly demolished the notion that science progresses in an orderly linear fashion, with one discovery leading inevitably to the next.

> The progress of Science is generally regarded as a kind of clean, rational advance along a straight ascending line; in fact it has followed a zig-zag course, at times almost more bewildering than the evolution of political thought. The history of cosmic theories, in particular, may without exaggeration be called a history of collective obsessions and controlled schizophrenias; and the manner in which some of the most important individual discoveries were arrived at reminds one more of a sleepwalker's performance than an electronic brain's (Koestler, 1963, p. 15).

Koestler exaggerates the irrational element in science, but even scientists themselves have little sympathy with traditional views of the scientist. As Agnew and Pyke (1969) note, "We suspect that there are those who . . . will say that the researcher must be completely dedicated to objectivity, that he is only interested in the truth. Perhaps there are researchers like that. We haven't met enough to fill a phone booth" (p. 162).

Scientists are not totally objective, dispassionate observers and science has not progressed in a completely rational manner. But there has been progress: modern scientific theories can predict and explain a much wider range of events than their predecessors of a hundred years ago. The modern theory of plate tectonics, for example, accounts for a mass of geological data that no previous theory could have handled (Gould, 1977).

Falsification as a Demarcation Between Science and Non-Science

The philosopher Karl Popper (1962, 1976) concedes that the way in which scientific discoveries are made is often non-rational. But he claims the way in which they are tested is rational--or at least potentially rational. Einstein's General Theory of Relativity is Popper's favorite example. When he proposed the theory, Einstein also proposed an empirical test that could refute it. He predicted that beams from a distant star would be bent a certain amount by the sun's gravitational field. During an eclipse, the British physicist A. S. Eddington found that light from a star near the sun was bent by the sun's gravity in the manner predicted by Einstein. General Relativity had not been disconfirmed by an empirical test.

Note that it is not correct to say that Einstein's theory had been confirmed by an empirical observation. That is the logical fallacy called 'affirming the consequent.' Another theory can always be constructed that will make the same prediction. In fact, other hypotheses besides Einstein's have been proposed that account for Eddington's observation (Kaufmann, 1973).

Restating the problem in logical terms will make it clearer. Scientific predictions are "if, then" statements: if hypothesis p is true, then event q will be observed.

There are only two forms of valid arguments involving one if, then statement and another, non-binary statement:

1. Modus ponens: Hypothesis p is true. Then event q will be observed.

2. Modus tollens: Event q is not observed when hypothesis p predicts it should be. Then p is false.

If q is observed, that says nothing about whether p is true. An alternative hypothesis h might also predict that q would have occurred.

Popper's point is that modus tollens is the only viable form of scientific inference for testing hypotheses. Even modus ponens isn't useful because a scientist cannot know a priori that a hypothesis is true. The logical conclusion is that a hypothesis can never be proved right, but it can withstand repeated attempts to prove it wrong using modus tollens.

Another line of reasoning that supports Popper's can be derived from Hume's critique of induction (Popper, 1962). Put simply, Hume showed that one can never infer truth about the future by observation, because one cannot assume the future will be like the past. "Bertrand Russell once speculated that the chicken on slaughter-day might reason that whenever the humans came it had been fed, so when the humans would come today it would also be fed. The chicken thought that the future would resemble the past, but it was dead wrong" (Skyrms, 1966, p. 27). In the same way, even though scientific equations have given accurate predictions in the past, we can never be absolutely sure they will continue to do so in the future.

Again, the point is that no amount of evidence will prove that a theory about the universe is right, but one contradictory piece of evidence can prove it wrong. So, despite Hume's critique, scientific progress is still possible. What scientists should do is develop theories that make falsifiable predictions and then try to disprove them. Those theories that are not falsified represent the closest approximations to the truth available. Eventually, they may be disproved and replaced by other, better theories that account for even more evidence. The classic example is Newtonian mechanics, which were scientific gospel until Einstein's General Relativity came along and accounted for some evidence--like the perturbations in Mercury's orbit--that Newton's theory could not explain (Einstein and Infeld, 1938).

How does one distinguish between General Relativity and its competitors, if some of the competitors make the same predictions as Einstein's theory? Popper says that the best theory is the one which forbids the most, i.e., makes the most potentially falsifiable predictions. Theories like Marxism and psychoanalysis that make no falsifiable predictions are not scientific. No matter what neurosis you bring to a psychoanalyst, he or she will always be able to explain it in Freudian--or Adlerian, or Jungian--terms. No matter what world event occurs, a Marxist will always be able to explain it in terms of the struggle between classes. These theories forbid nothing: they make no predictions that can be disproved. Their explanations are deceptive: they can explain what has happened, but they cannot make specific, falsifiable predictions about future events, as Einstein did.

So, according to Popper, the demarcation between science and non-science is that all scientific theories are falsifiable: they make predictions that potentially can be disproved. Science approaches the truth by discarding ideas that are wrong, not by proving ideas right. Even Einstein's theories may someday be supplanted by better approximations to the truth.

The fact that many scientific ideas are the products of intuition rather than reason does not bother Popper. It does not matter where a scientific idea comes from: it matters only that it make falsifiable predictions. If the idea is ridiculous, it will be disproved immediately.

Problems with Popper's Demarcation

Historians of science like Kuhn (1970) have shown not only that individual scientific ideas often arise from irrational sources but also that scientific theories are not always accepted or rejected based on the science. For example, Einstein's special theory of relativity was initially falsified, in an experiment by the eminent physicist Walter Kaufmann. His results supported theories that differed from Einstein's. Einstein was not at all dismayed; as far as he was concerned, the ad hoc character of the other theories rendered them highly unlikely (Holton, 1973). About ten years later, it was discovered that Kaufmann's equipment had been inadequate to conduct the test. Einstein had refused to let a single experiment sway his opinion, even though he had agreed that--at the time at least--there were no problems with the way the experiment had been done.

Popper would respond by arguing that falsification is an ideal. Whether scientists have practiced it in the past or not is irrelevant. The important thing is that they should practice falsification in the future.

But the Einstein example raises problems with falsification as an ideal. Einstein was correct to stick to his theory despite evidence to the contrary. Sometimes, falsificatory evidence should be ignored.

Also, even if we accept falsification as a scientific ideal, it does not serve to demarcate science from non-science. There are major scientific areas that are based on non-falsifiable theories. Popper himself admits that "Darwinism is not a testable scientific theory, but a *metaphysical research programme*--a possible framework for testable scientific theories" (1976, p. 168). Like Marxism, Darwinism can be used to explain any event within its theoretical domain--any new discovery in the fossil record, any new and strange form of life. Evolution cannot be used to predict the specific course of an organism's development, but it can account for the past development of any organism.

According to Popper, the only major prediction made by

Darwin's theory is that changes produced by selection working on variation will be gradual. Small variations in an organism are selected by the environment until, over many generations, the species is transformed. This prediction has not been falsified by anything in the fossil record. But it is hard to derive other testable predictions from evolutionary theory.

Much of modern biology is based on evolution. Is biology a science? Not if its major theoretical framework is not falsifiable. Popper's demarcation excludes biology from the sciences.

There are theories outside of science that fulfill Popper's falsification criterion. Tradition held that the first five books of the Bible were written by Moses. This idea can be translated into a falsifiable hypothesis: if the first five books exhibit strong stylistic and thematic inconsistencies that suggest they were written by several authors, they could not have been written by Moses. In fact, modern Biblical scholars have used this kind of reasoning to prove that Moses could not have written the first five books and that at least four different authors are responsible (see The Oxford Annotated Bible, 1962, p. xxiv). Is Biblical scholarship a science?

How about history? Good historical theories can be falsified. Take, for example, the common notion that Einstein's theory of special relativity was inspired, in large part, by the Michelson-Morley experiment. Gerald Holton

(1973) showed that Einstein was not even aware of this experiment at the time he wrote his famous 1905 paper.

Popper has proposed an ideal: every scientific theory should be open to disconfirmation. But why limit the ideal to science? Fields that are not traditionally scientific can also generate falsifiable hypotheses.

Should scientists--and other scholars--concentrate on falsifying their theories, rather than verifying them? Recently, a few researchers have begun to explore this question--using the scientific method.

Using Science to Study Confirmatory Reasoning

In the preface to his book Scientist as Subject: The Psychological Imperative, Michael J. Mahoney (1976) noted that, "Relative to the last century--or even the last decade --today's scientists know quite a bit about virtually everythnig on our planet--with one ironic exception. The exception, of course, *is* the scientist" (p. xi). According to him, historians and sociologists have given us mainly "biographies and social systems analyses so that we are still left with a very meager understanding of the *psychology* of the scientist" (p. xii).

Mahoney and others have recently set out to remedy this defect by doing experiments and surveys designed to increase our understanding of how scientists reason. The focus of these studies has been on whether scientists--and ordinary college students--can successfully employ falsification on problems that model scientific logic.

Mahoney and Kimper (1976) gave a questionnaire to 400 scientists which included a test of their logical reasoning skills. The scientists were asked to pick which of four types of inference were valid and were given a specific example to see whether they could use each type correctly. They were told to assume "if p, then q," then asked which of the following conclusions were valid:

Observation	Conclusion
p	q
not-p	not-q
q	p
not-q	not-p

Only the first (modus ponens) and fourth (modus tollens) are valid. The second is the fallacy called 'affirming the consequent' and the third is the fallacy called 'denying the antecedent.'

The specific example asked scientists to "Assume that the four boxes which are presented below are actually cards which each have a letter on one side and a number on the other side. You are asked to test the hypothesis that--for these 4 cards--if a vowel appears on one side, then an even number will appear on the other side. Your "testing," of course, will involve turning one or more cards over" (Mahoney, 1976, p. 189). The first box had an *e* in it, the second an *m*, the third an *8* and the fourth a *7*. The two critical cards are one (modus ponens) and four (modus tollens).

Eighty-two scientists returned the survey. They included physicists, biologists, psychologists and sociologists. "Fewer than eight per cent of the scientists were able to identify the irrelevant 'experiments' in the analogue hypothesis-testing task. Fewer than ten per cent correctly selected the experiments which had the critical potential of falsifying the sample hypothesis. Likewise, although the vast majority of subjects were able to recognize the validity of modus ponens (confirmation) and the invalidity of affirming the consequent, almost 30 per cent of the social scientists incorrectly rated denying the antecedent as a valid form of reasoning. Most interesting, perhaps, is the finding that over half of the scientists did not recognize modus tollens (disconfirmation) as being logically valid" (Mahoney, 1976, pp. 192-193).

Mahoney could not guarantee that his sample was representative, because not all the scientists he contacted returned his survey. But of the ones who did, over half did not recognize the logical validity of falsification and even more had trouble applying it in a simple example.

Kern, Mirels and Hinshaw (Note 1) gave a questionnaire similar to Mahoney's to a sample of seventy-two faculty members--psychologists, biologists and physicists--from a large midwestern university. Half the sample were given logical problems written in abstract form (if p, then q) and half in concrete form (if Rex is a terrier, then he likes apples). Kern et al. used the following problem to express modus tollens in concrete form (Kern, Mirels & Hinshaw, 1980):

"Given: If Rex is a terrier, then he likes apples.

Observation: Rex does not like apples.

Conclusion: a) Rex is not a terrier.

b) Rex is a terrier.

c) Rex likes apples.

d) None of the conclusions seem to follow logically" (p. 23).

Thirty per cent of the scientists surveyed did not select the proper response (a) to the above question, indicating that they could not correctly apply modus tollens to a specific example. Half the scientists presented with modus tollens in its abstract form failed to recognize its logical validity, replicating Mahoney's findings. Kern et al.'s sample was somewhat more representative: all those contacted responded, but all the scientists were from the same university.

The questions used in these surveys are not perfectly analogous to the research situations faced by working scientists. But it is still surprising that so many respondents failed to make proper use of falsification, even when given concrete problems.

Another replication of this pattern of results is provided by Einhorn and Hogarth (1978) who used a sample of subjects "known to have been trained in examining possible disconfirming evidence" (p. 399)--twenty-three statisticians at the University of London. Statisticians are formally trained in hypothesis-testing, including conditions under which a hypothesis should be rejected.

Einhorn and Hogarth asked statisticians to indicate which of four outcomes could be used to evaluate a claim: "when a particular consultant says the market will rise (i.e., a favorable report) it always does rise" (Einhorn & Hogarth, p. 399). The four outcomes were:

"1. favorable report.
2. unfavorable report.
3. rise in the market.
4. fall in the market" (p. 399).

Only five of the statisticians indicated both critical outcomes: one (confirmatory) and four (disconfirmatory). Twelve selected the confirmatory outcome alone. Like Mahoney's scientists, most of the statisticians recognized the validity of modus ponens but not modus tollens.

The confirmatory bias shown by scientists and statisticians is shared by college students. Wason (1977) gave students a numerical triad (2, 4, 6) and told them the triad conformed to a simple mathematical rule. To discover the rule, subjects were allowed to try as many other three-number strings as they wanted, and the experimenter told them whether each was right or wrong. Subjects were run singly. When each was confident that he or she had discovered the rule, he or she told the experimenter.

Out of twenty-nine subjects, twenty-two "announced at least one incorrect rule, nine of these announced a second incorrect rule, and two of these nine announced a third incorrect rule. Six subjects announced the correct rule without any incorrect ones, and the results showed that these subjects varied their hypotheses much more frequently than those who announced one incorrect rule" (Wason, 1977, pp. 308-309). Most of the students tried to verify their guesses, instead of falsifying them. If a student decided that numbers in a triad must go up by twos, he or she would often announce several strings like 8, 10, 12 and 4, 6, 8 to test that idea. When the experimenter confirmed that these strings were correct, the student often concluded that his or her guess was correct--even though many alternatives had not been explored. The few subjects who got the correct answer without any incorrect guesses varied their hypotheses more than the other subjects. In addition, they used strings that would tend to disconfirm their guesses, e.g., 2, 3, 4, to test whether a difference of two is really necessary. Finally, these subjects were more conservative about making a guess; they waited until they had thoroughly tested their hypotheses before announcing them to the experimenter (see Wason, 1977, pp. 309-310 for some examples of subjects' strategies).

Wason's study showed both that college students prefer confirmatory evidence and that those college students who tried to disconfirm their ideas did better on a simple problem-solving task.

Mahoney and DeMonbreun (1975; reported in Mahoney, 1976) used Wason's task on a different subject population: psychologists, physical scientists and Protestant ministers.

They chose fifteen representatives from each group. The only two errorless subjects were ministers. Scientists used slightly fewer disconfirmatory trials and tended to be more speculative: they generated more hypotheses than the ministers and returned to a previously falsified hypothesis more often. This result contradicts the traditional image of the scientist as more conservative and cautious about proposing new hypotheses. The small sample-size in Mahoney's study raises questions about whether it generalizes to the larger population of scientists, but the results are provocative, nonetheless.

Some support for Mahoney's conclusions comes from another, nonexperimental study. Mitroff (1974) observed the reactions of forty-two geoscientists to the Apollo missions. Many had committed themselves to major hypotheses concerning what Apollo would find. Mitroff interviewed them at various times across the course of the Apollo missions. He found that some of these scientists tended to seek evidence confirming their hypotheses and discredited contradictory evidence. Like Mahoney's subjects, they stuck to their hypotheses even after it became apparent that they had been falsified. Moreover, a number of scientists argued that this kind of commitment and bias in science was a good thing. As noted earlier, initial experimental evidence disconfirmed Einstein's special theory of relativity but he correctly dismissed the evidence.

An elegant study with introductory psychology students provides added evidence for a confirmatory bias on the part of most individuals involved in scientific problem-solving. Mynatt, Doherty and Tweney (1977) created a scientific 'environment' using a computer terminal which displayed three shapes: triangles, squares and discs. Each shape was either completely lit or half lit.

Forty-five subjects were seated individually at terminals and told how they could make a small lighted dot or 'particle' move across the screen, through a field which contained a specific arrangement of triangles, squares and discs. Each subject was supposed to come up with a hypothesis that would account for the motion of the particles. They were given one of three kinds of instructions: disconfirmatory, confirmatory, or instructions that merely told them they should test their ideas. The confirmatory and disconfirmatory instructions included a historical example of each.

The only rule governing the motion of particles was that when a particle came within four centimeters of the center of a half-lit figure, it stopped. But the environments the subjects faced were arranged so that it appeared triangles might play some role in stopping motion: all triangles that appeared in the first two screens the subjects faced were either half-lit or within the boundary of a half-lit figure. As a result, twenty subjects formed initial triangle hypotheses, while only twelve had hypotheses concerning brightness and thirteen had various other ideas.

After their first two screens, subjects were showed photographs of other screens--including different arrangements of shapes and brightness--and asked which ones they would like to use to test their hypotheses. Some screens were designed so that they would provide confirmatory evidence for a triangle hypothesis; others were arranged so that it would be easy to disconfirm such a hypothesis. Of the twenty subjects who decided initially on a triangle hypothesis, fifteen initially chose confirmatory screens. This tendency continued: the twenty 'triangle' subjects selected confirmatory screens on over seventy per cent of their choices.

Then subjects were given the opportunity to fire particles at either the screens they had chosen or ones they had not. Eleven of the 'triangle' subjects obtained evidence falsifying their hypothesis and ten of these achieved the correct solution. Of the other nine, only four were ultimately correct.

The instructions to either confirm, disconfirm or test had no effect on whether individuals arrived at the correct solution. The only thing that made a difference was the initial hypothesis. If subjects focused on brightness right away, they got it. If they focused on something else, they did considerably worse.

This study supports Wason's observation that introductory psychology students, like scientists, prefer to confirm hypotheses, rather than falsify them. In Mynatt et al.'s study, once subjects obtained evidence falsifying a theory, they were able to use it. But in Mahoney's study with

DeMonbreum (using Wason's task) scientists showed a persistent tendency to return to their initial hypotheses, despite falsifying evidence. Are introductory psychology students better able to evaluate disconfirmatory evidence than working scientists? Two studies provide too slim a basis for such a generalization--especially when the studies use entirely different tasks. Nonetheless, it is an intriguing finding, and might be explained by the fact that norms governing publication and advancement in most sciences reward those whose hypotheses are confirmed. A study by Spencer, Hartnett and Mahoney (Note 2) showed that journal referees in psychology preferred confirmatory results to disconfirmatory ones. Could scientists be systematically biased towards confirmation?

All of the scientific problem-solving studies cited above concluded that there is a significant advantage to disconfirmation, even though only two of the studies, Wason's and Mynatt et al.'s, demonstrated the advantage on an actual problem-solving task. Only one study--Mynatt's--actually involved training subjects to use disconfirmatory reasoning and the training manipulation in that study produced no effects. Presumably, subjects ignored it. Clearly, there is a need for further research in this area. If disconfirmation is the best way to test theories, then one ought to be able to train subjects to falsify and see that training work when subjects tackle problems that model scientific reasoning.

Should Groups Falsify?

All the experimental studies of scientific problem-solving have focused on whether individuals working separately can effectively employ a disconfirmatory strategy. But some of the most significant advances in science have been made by groups of scientists working together, including the development of atomic energy (Sherwin, 1973) and the discovery of DNA (Judson, 1979). There are scientists like Einstein who do much of their work alone, but even these 'loners' communicate constantly with colleagues through journals, conferences and correspondence.

Is falsification a practical strategy for every member of a scientific community? A number of the scientists studied by Mitroff (1974) argued that researchers should be committed to their hypotheses and defend them against disproof. Einstein did just that with his theory of special relativity and his theory was correct, not the experimental evidence. As A. S. Eddington remarked, "It is...a good rule not to put too much confidence in the observational results that are put forward until they are confirmed by theory" (quoted in Judson, 1979, p. 93). There are times when theory should take precedence over empirical evidence.

If one scientist argues vehemently for his or her hypotheses,others will have to provide powerful disconfirmatory evidence before the theory will be rejected by the scientific community. Rather than having every researcher adopt a disconfirmatory strategy, it may be better for science to operate on this kind of an adversary system, each theory having its proponents who try to confirm it and its opponents who try to disprove it.

At this time, it is impossible to say which approach is best for science. But it should be apparent that a strategy that works well for isolated individuals may not work well for groups. There is a need for research comparing how groups use confirmatory and disconfirmatory reasoning.

Unfortunately, none of the studies of scientific reasoning have included groups. But there is an extensive literature concerning the advantages and disadvantages of problem-solving in groups.

Group Problem-Solving

Gerald Holton argues that "the contributions of n really good persons working in related areas of the same field are likely to be larger (or better) than n times the contribution of any one of them alone in the field. This is true of a group as well as of individuals who do not work in physical proximity to one another" (1973, p. 409). So, a group of four scientists working in the same area would make more than four times the contribution of a single scientist working alone.

Holton's idea is not entirely supported by experimental research on group problem-solving, even though none of the group problem-solving studies have used tasks designed to model scientific reasoning. Shaw (1932) compared the performance of four-person groups and individuals working separately on the following problem:

> On one side of a river are three wives and three husbands. All of the men but none of the women can row. Get them all across the river by means of a boat carrying only three at one time. No man will allow his wife to be in the presence of another man unless he is also there (Steiner, 1972, p. 19).

Sixty per cent of the groups solved the problem, as opposed to only fourteen per cent of the individuals. In other words, groups were almost exactly four times as successful as individuals. Given that there were four people in each group, Shaw's finding can be interpreted to mean that groups do no better than the _best_ of an _equal_ number of individuals working separately. Subsequent research on tasks similar to Shaw's supports this conclusion (Marquart, 1955; Steiner, 1972).

Type of task is important. Groups will do better than an equal number of separate individuals on tasks that are divisible into sub-tasks, if each group member has special skills relevant to one or more sub-tasks (Steiner, 1972). The team of scientists that developed the atomic bomb is a good example: chemists, physicists, engineers and politicians handled different aspects of this highly complicated, divisible task (Sherwin, 1973). No one scientists or technician could have come close to solving this problem.

The tasks used by Shaw (1932), Marquart (1955) and others --on which groups did only as well as the best of an equal number of individuals working separately--are not divisible. They are what Steiner (1972) calls 'Eureka' problems: the solution comes to a single person in a 'flash' of insight.

Lorge and Solomon created a mathematical model to predict group performance on a Eureka task. "If, in a given population, the proportion of people not possessing the ability . . . is Q, the probability of drawing at random from the population a single person who will not have the ability to solve the problem is Q. The probability that nobody in a randomly assigned group of size n will have the ability to solve the problem is Q^n, and the probability that at least one member of the group will be able to solve it is $1-Q^n$. Thus, if the presence of at least one competent member is sufficient to guarantee group success, the proportion of successful groups will equal $1-Q^n$" (Steiner, 1972, p. 20).

In Shaw's study, the proportion of individuals who solved the problem working alone was .14, so P=.14 and Q=.86. Since there were four members in each of Shaw's groups, the probability that at least one member of each group would solve the problem was $1-(.14)^4$, or .596. The actual proportion of groups that solved was .6, so the Lorge-Solomon model is a good predictor in this case. The model has been successfully applied to other, similar studies (Steiner, 1972).

Holton's remark that "the contributions of n really good persons working in related areas of the same field are likely to be larger (or better) than n times the contribution of any one of them alone in the field" would seem to hold true for divisible tasks, but not for 'Eureka' problems that can be solved by one individual's insight. A good example of such a Eureka insight is Einstein's special theory of relativity. He exposed some serious logical inconsistencies in the physics of his day and resolved them by means of a dramatic new theory (Holton, 1973). One of the paradoxes that absorbed Einstein was how a beam of light would look if he could run as fast as it. It had never occurred to anyone else to puzzle over such a problem. This kind of imagination was critical to the development of the special theory of relativity.

Therefore, it appears that scientific problems which require a kind of 'Eureka' insight can be solved at least as easily by a talented individual as by a group. Scientific problems that can be divided into sub-tasks are much easier for groups to solve--because each member of the group can focus on part of the problem. But these conclusions should be subjected to further experimental test, because none of the group problem-solving studies have used tasks specifically designed to model scientific reasoning and/or have included scientists as subjects.

The Present Study

So far, we have discussed two separate experimental research traditions: falsification by individuals, and group problem-solving. The present study is an attempt to combine elements of both in a single experiment.

Why an experiment? The advantage of using a laboratory simulation of science is that one can manipulate variables like type of strategy while eliminating the effects of other variables via control procedures. It is hard to disentangle the effects of one variable from another in the 'real world'; proper laboratory procedures permit a clear test of how, for example, confirmatory and disconfirmatory strategies affect performance on a particular problem.

The disadvantage of experiments is that laboratory simulations are not perfect models of the things they are designed to simulate. One usually sacrifices realism for control in an experiment.

The obvious solution to this dilemma is to tackle a problem using multiple methods--experiments, surveys, biographies, historical studies, interviews, etc. Since there are few experimental studies of scientific problem-solving, and no studies concerning how groups solve scientific problems, I chose to do an experiment involving groups and a task that models scientific reasoning.

To combine the two research traditions discussed earlier, the experiment presented here involved manipulation of three variables.

(1) Strategy: To see whether falsification is an effective strategy for individuals working in communication with one another, groups of four subjects were told to adopt either a confirmatory or a disconfirmatory strategy on a task that models scientific reasoning.

(2) Communication: To see whether individuals working together would do better than individuals working separately, subjects were run in either of two kinds of groups. Members

of interacting groups worked together and discussed ideas freely. Members of co-acting groups worked separately and kept their guesses to themselves.

So, each group of four subjects was assigned to one of four conditions: 1. an interacting group told to use a disconfirmatory strategy; 2. an interacting group told to use a confirmatory strategy; 3. co-acting/disconfirmatory; 4. co-acting/confirmatory. All subjects were run under the same conditions on the third variable.

(3) Task: The task used in the present study had to fulfill four requirements:

(a) It had to be a good model of scientific reasoning, particularly of the Eureka-type, because most previous studies of scientific problem-solving had used Eureka problems.

(b) It had to be simple, so subjects could potentially solve it in the short time they would be together in the laboratory.

(c) It had to be divisible into a sequence of related subtasks that increased in difficulty, so groups could gain experience working together on simpler problems before tackling harder ones.

(d) Like Wason's and Mynatt et al.'s tasks, the one used in the present study had to permit subjects to perform experiments to test their guesses. Subjects should be able to design both confirmatory and disconfirmatory experiments.

Fortunately, such a task exists.

New Eleusis

The card game "New Eleusis," as described by Martin Gardner in one of his Scientific American columns (October, 1977), is explicitly designed to model the "search for truth." "...Eleusis is of special interest to mathematicians and scientists because it provides a model of induction, the process at the very heart of the scientific method... Eleusis was invented in 1956 by Robert Abbott...He had been studying that sudden insight into the solution of a problem that psychologists sometimes call the 'Aha' reaction. Great turning points in science often hinge on these mysterious intuitive leaps. Eleusis turns out to be a fascinating simulation of this facet of science" (Gardner, 1977, p. 18).

Here is a task that fulfills requirement (a): it models the Eureka aspect of scientific reasoning. Eleusis is usually played by four or five people. The dealer makes up a rule that determines when a card will be right and when it will be wrong. Rules can be simple or complex, straightforward or ambiguous; therefore, Eleusis fulfills requirement (b). The other players take turns playing cards, with the dealer telling them which are right and which are wrong. Correct cards continue in a straight line; incorrect cards are placed at right angles, under the card they followed.

For example, if the rule is "red and black cards must alternate," the sequence might end up looking something like this after a few cards had been played: (H=hearts, D= diamonds, S=spades, Q=queen)

10H	5C	2D	7S	4H	3C	9D	QC
	9S			7D			6S
	QC			QD			

Players can always see the results of each others' past card plays, or experiments.

Requirement (c) was fulfilled by designing a set of four relatively simple rules that increased in difficulty, with solutions to later, more complex rules building off solutions to the earlier, simpler ones. In this way, interacting groups, like scientific research teams, could gain experience working together, starting with simpler problems and building to more complex ones.

Martin Gardner claims that Eleusis "provides a model of induction, the process at the very heart of the scientific method." But Hume and Popper have raised serious question about induction: it is not a truly valid form of inference and most scientists don't really use it (Popper, 1962). Scientists don't merely 'observe': they look at the data in the light of hypotheses they have already formed.

If Eleusis is a good model of induction, then it is not a good model of the ideal scientific procedure, as Popper sees it. The game is structured so that players are rewarded for getting cards right, which is why it is best played using a kind of inductive, confirmatory strategy: if your idea for a rule generates right answers, keep playing it. In most sciences, those whose experiments confirm their hypotheses are much more likely to get published and advance in their field (Spencer, Hartnett & Mahoney, 1980). So the emphasis on right answers in Eleusis may be a realistic simulation of the way science is conducted, but it does not conform to the Popperian ideal.

Fortunately, it is possible to transform Eleusis so that players will not be rewarded or punished for following any particular strategy--or even rewarded for following a disconfirmatory strategy. Eleusis makes a very flexible task: with only a few modifications, it can be changed from a model of induction to a model of falsification by requiring players to write out their hypotheses, then test them by playing cards that should be wrong. In this way, Eleusis can be made to fulfill requirement (d).

Speculations Concerning the Results

Interacting groups could potentially pursue a disconfirmatory strategy more consistently than co-acting groups, because interacting group members could learn to coordinate efforts on the earlier, easier rules, so that by the later rules, they would be falsifying systematically. In co-acting groups, each individual would be more likely to pursue his or her own strategy, independently of other group members.

So, if Popper's ideas concerning the advantages of falsification are correct, then interacting groups trained to falsify have the potential to do better than co-acting groups. Conversely, interacting groups trained to confirm should do worse than co-acting, because group members will be consistently pursuing a disadvantageous strategy.

Of course, all of this assumes that interacting group members will be able to coordinate efforts. The literature on group problem-solving indicates that there are many situations in which the processes by which an interacting group arrives at a decision are so inefficient that the group performs much worse than one would expect (Steiner, 1972; Hoffman, 1979).

Three Comparisons Between Interacting and Co-Acting Groups

There are three comparisons that should be made across the intragroup communications variable.

Interacting groups vs. the best co-actor on each rule. Will the best individual in each co-acting group on each sub-task perform better than the interacting groups? Marquart (1955) found that on a single task, there was no difference in performance between an interacting group and the best of an equal number of individuals. So if we compare performance, sub-task by sub-task, looking to see if any member of each co-acting group solved each rule, there should be no difference between interacting and co-acting groups.

Interacting groups vs. the best co-actor across the four rules. Will interacting groups perform better than the best individual in each co-acting group when best is defined as the person who solves the most rules? Steiner (1972) argues that, on a task that is divisible into sub-tasks, interacting groups will perform better than an equal number of individuals working separately. The four Eleusis rules used in the present study can be viewed as parts of a single, divisible task because the rules are related: solving each rule makes it easier to solve the next. If we take the member of each co-acting group who does the best across all four rules, his or her performance should be worse than that of an interacting group--because in an interacting group, one person can solve the first rule, share his or her solution with the others and make it possible for someone else to solve the second rule.

Do interacting groups solve a higher percentage of problems than co-actors? This comparison comes closest to Marjorie Shaw's (1932) observation that groups solved a higher percentage of problems than individuals working separately. In the present study, Shaw's finding should certainly be replicated: the proportion of successful interacting groups should be much greater than the proportion of successful co-acting individuals. But the co-acting individuals are not working in complete isolation. They do share a certain amount of information with each other. So the results of the present study might differ from Shaw's.

One other dependent variable that is commonly studied in the literature is time to solution. Individuals working separately are almost always faster than interacting groups (Kelley & Thibaut, 1969). This relationship should hold in the present study: those co-acting individuals who solve the problem should take less time than groups working together. Co-actors should also take less time to make a decision concerning when to give up and admit they can't solve the problem. Groups are rarely efficient in terms of time. Their main advantage--when they have one at all--is in terms of quality of solution.

How the Present Study Relates to the Practice of Science

The laboratory simulation of scientific problem-solving presented here is a very simplified model of scientific processes in the real world. The interacting/co-acting comparison is analogous to a comparison between research teams that work together and scientists that see only the results of each others' experiments in journals. The confirmatory/disconfirmatory comparison is analogous to a comparison between scientists who seek to verify hypotheses and scientists who try to disprove them. But the analogy is weak--especially as the subjects in the present study will be introductory psychology students, not scientists.

The game "New Eleusis" is a good model of scientific problem-solving (Romesburg, 1979). But the four rules used in the present study are much simpler than most real scientific problems. Eleusis rules of great complexity could be developed for use in future studies, but it seemed better to begin with relatively simple ones that most college students could potentially solve in a short time.

So, the results of the present study can only be generalized to the process of scientific discovery with the greatest caution. The laboratory is a long way from reality.

But this study is intended as only the first, exploratory step towards more realistic experimental simulations of science. Eventually, scientists themselves can be brought into the laboratory and asked to solve Eleusis rules of enormous complexity. Simulations of the growth and development of science could be set-up, using college students. The possibilities are endless. The main purpose of the present study will be to suggest fruitful lines of inquiry for future research.

II. Methods

Subjects

One-hundred-and-seventy-two introductory psychology students at the University of New Hampshire participated in this experiment. They were run in groups of four. An effort was made to insure that there were two males and two females in each group. On rare occasions, three members of one sex and one of another had to be run. All subjects were randomly assigned to one of the four possible combinations of the Strategy and Communication variables.

Data from three groups (twelve subjects) had to be dropped because of procedural errors made by the experimenter.

Procedures

Each subject was handed one of four cards marked "I," "J," "K," or "L" before he or she entered the experimental room. The letters corresponded to seats around a rectangular table. Same-sex pairs were handed either I and K or J and L to insure that they faced each other.

Once all subjects were seated, the experimenter handed-out written instructions and read them aloud. The first sheet of instructions concerned the task subjects were going to have to perform. The game "New Eleusis," discussed in the last section, was modified to fit the needs of the present experiment. Subjects were told that they would be dealt a hand of thirteen cards from a shuffled deck. Each person would then play a card, in order. After a card was laid on the table, the experimenter would indicate whether it was right or wrong and put it in the appropriate place in the sequence. Correct cards would continue in a straight line; incorrect cards would go off at right angles.

The experimenter then showed subjects an example of a simple rule: "red and black cards must alternate." Subjects were also made aware of the numerical values of the cards: Ace is a one, Jack is an eleven, Queen is a twelve and King is a thirteen.

The first sheet of instructions concluded by pointing out that there would be four separate tasks. Subjects would play sixty cards on each and be limited to half-an-hour in which to guess the rule. To make sure subjects always had a card to play, each time a person played a card that was wrong, he or she would be given two additional cards. It was emphasized that this was not a penalty or punishment of any kind; in fact, the more cards one had, the better off one was.

The next sheet of instructions gave subjects specific directions for writing-out their guesses, including how to put down the time and card number on which they made their guess. When the group or an individual got ready to make a guess, they told the experimenter, who read them the time and made sure they had the correct card number. Time was measured in five-second intervals on a Hewlett-Packard 67 calculator, which has a timing program. The time provided a measure of speed to solution, while the number of cards provided a rough measure of the amount of information available as of each guess. Subjects were also told they would not get any feedback on whether their guesses were right or wrong until the end of the experiment.

Independent Variables

After the procedure for making guesses was explained to subjects, including the fact that they could make as many guesses as they wanted, subjects were either told to make their guesses as a group (interacting condition) or keep track of their guesses separately (co-acting condition). In the former case, it was emphasized that subjects should discuss their ideas freely and even make suggestions concerning what card should be played next. In the latter case, it was emphasized that subjects should not say anything to one another about their guesses or ideas.

Next, subjects were given a final sheet which contained a list of "Good Strategies for Guessing Patterns." First, all groups were urged to "guess early, guess often." It was hoped that this suggestion would encourage subjects to leave a record of all their ideas on each task.

Second, half the groups were urged to "systematically test your guesses by looking carefully at previous mistakes and by playing cards you are sure will be wrong." The other half were told nothing about mistakes and urged to test their guesses by playing cards they were sure would be right. The former strategy is disconfirmatory, the latter confirmatory.

Upon completing the instruction, subjects began on the four rules. In the order in which subjects saw them, they were:

1. The next card must be one higher or lower than the preceding card.

2. The next card must be the same as, one higher or lower than, or two higher or lower than the preceding card.

3. Odd and even cards must alternate: an even card can never follow an even card and an odd card can never follow an odd card.

4. Cards can alternate odd-even, or red-black, or both. If an even card follows an even card of the same color, or an odd card follows an odd card of the same color, it is wrong. Any other two-card combination is possible.

The first three rules all concern the numerical values of the cards and can be solved by focusing on the differences between two adjacent cards. Even the third rule can be stated in terms of differences: if the difference between two adjacent cards is odd, one will be odd and the other will be even.

The fourth rule combines two dimensions--color and number--and can be solved only by attending to both of them at the same time. But both dimensions are ones the subjects have encountered previously. The odd-even dimension formed the basis for the preceding rule, and alternating colors was used in the initial instructions as an example of how Eleusis is played.

If subjects adopt a purely confirmatory strategy, they can generate right answers to all four rules by playing according to the first rule: cards must be separated by a difference of one. Information derived solely from right answers will be misleading. To guess the four rules correctly, subjects will have to make use of information derived from mistakes. Hence, a disconfirmatory strategy should be superior to a confirmatory one.

Dependent Measures

The term dependent is used to refer to the class of variables that are dependent on subjects' responses. Four measures were used to assess subjects' performance:

1. The guesses: For purposes of analysis, correct guesses were coded as ones and incorrect guesses are zeroes. A record was kept of all *individual* guesses in the co-acting condition and all *group* guesses in the interacting condition.
2. Number of cards: Each time subjects wrote down a guess, they also wrote down the number of cards that had been played on that rule as of the point where they made the guess. Therefore, card totals could be anywhere between one and sixty.
3. Time to solution: Every time subjects made a guess, they also wrote down the time, in seconds, that had elapsed since the beginning of that rule. Separate times were therefore available for each of the four rules.
4. Mistakes: The total number of cards each group got wrong on each rule was noted. This record provided a manipulation-check for the strategy condition: disconfirmatory groups should make more mistakes than confirmatory.

In addition to the performance measures, all groups were tape-recorded to provide information about group processes.[1]

After subjects had completed the four tasks, they were given a questionnaire which asked them about their previous experience with card games, the number of science and math courses they had taken in college, their reactions to the task and whether they had found the suggestions concerning strategies helpful. Information regarding their major, their class in college (freshman, sophomore, etc.) and whether they had a close friend in the group was also obtained.

At the end of the experiment, subjects were debriefed and sworn to secrecy. The debriefing included a review of the procedures, the correct solutions to the rules and the rationale behind the design. It should be emphasized that this was the only time subjects were given any feedback on their guesses. After the debriefing, subjects were thanked for their participation.

Overview of the Design

The design is what Winer (1971) calls a three factor experiment with repeated-measures on one factor. There are two between-group factors: type of strategy and intragroup communication. The task is a within-group factor: all groups get the same task and within each group the task is varied in the same way--by splitting it into four rules. The fact

[1] Unfortunately, due to a lack of experimental assistants, tapes could not be coded so as to provide quantitative information concerning the kinds of remarks made by specific individuals.

that subjects' performance on the task is assessed four times makes task a repeated-measures factor.

So, the design is a 2 X 2 X 4, with 2 (interacting or co-acting) by 2 (confirmatory or disconfirmatory) levels of the two between-group factors and 4 (the four Eleusis rules) levels of one within-group factor.

III. Results

Each of the three major dependent measures was analyzed using a 2 X 2 X 4 split-plot analysis-of-variance. This technique is described in Winer (1971, pp. 559-571) and Kirk (1968, pp. 283-294).

Questionnaire data was analyzed via a series of linear regressions.

Interacting Groups Vs. the Best Co-Actor on Each Rule

The first ANOVA compares the number of correct solutions obtained by interacting groups to the number of correct solutions obtained by the best co-acting individual on each rule, where best is defined as the person who solves each rule using the least cards and taking the least time. If only one co-actor solved each rule in a particular group, the group's score was all ones--even if a different person was successful on each task. This comparison is most rigorous, in that it demands interacting groups perform better than any co-acting individual.

The ANOVA table:

Source	SS	df	F
C	.10	1	.46
S	.02	1	.11
CS	.40	1	1.8
Er	7.95	36	

Source	SS	df	F	
T	2.87	3	6.21*	p<.0001
CT	.30	3	.65	
ST	.47	3	1.03	
CST	.20	3	.43	
Er	16.65	108		

The only significant effect on this comparison is a main-effect for task. Neither the strategy nor communication variables significantly affected the number of correct solutions a group obtained.

A look at the actual number of successful solutions (where ten is the highest possible number) shows the task effect clearly:

	Rule	1	2	3	4
Interacting	Disconfirmatory	9	8	7	6
	Confirmatory	9	9	9	6
Co-acting	Disconfirmatory	8	10	7	7
	Confirmatory	8	9	7	3
		34	36	30	22

Reading horizontally, one can see that the number of successes decreases as rule increases. But, except on task four, the number of successes are relatively equal across the other two factors. Only on task four is there any appearance of a difference due to either the communication or strategy variables. In the co-acting confirmatory condition, only three groups solved the fourth rule. But a separate ANOVA of solutions on rule four revealed no significant effects (see Appendix for source table). This means that we cannot be sure that the difference between the co-acting confirmatory groups and other groups on rule four wasn't simply due to chance.

There is one major problem with assessing the effects of the task in the present study. The fact that performance declined across the four rules could be due either to fatigue or an increase in rule-difficulty. Given that the effects of fatigue and practice might be expected to cancel each other out, however, it is likely that the main effect for task reflects an actual increase in the difficulty of the rules.

Analyses were also done on the number of cards and the amount of time (in seconds) it took successful groups to reach a solution. The number of cards is a rough measure of the amount of information each of the successful interacting groups and best co-acting individuals needed. This ANOVA was conducted in a slightly different way than the standard repeated-measures approach we have talked about all along. A repeated-measures analysis is not appropriate because most groups solved less than four rules, and consequently did not have scores on all four levels of the repeated measure. So a standard between-groups analysis was done, using the strategy and communication factors and treating the number of cards each group required to solve each rule as independent scores. There were potentially forty such scores in each cell (four per group) but in fact each cell had less.

The two independent variables had no effect on the number of cards required to reach a solution (see Appendix for source table). Those disconfirmatory groups that solved rules required no more information than confirmatory; similarly interacting groups required just as much information as co-acting groups. But remember, amount of information is also dependent on the kind of cards that have been played. One card that disproves a false theory can be worth thirty cards that appear to confirm it.

A similar analysis was performed using time-in-seconds as the dependent variable. Again, each individual time-to-solution was treated as a separate score, not as one of a series of four repeated-measures. The analysis turned up a significant main-effect for the communication variable:

Source	SS	df	F	
C	674961	1	5.12*	$p<.05$
S	1210	1	.009	
CS	145	1	.001	
Er	156837484	119		

The tiny F-ratios for the Strategy main-effect and the interaction indicate that the time data may not be normally distributed. So, both time and card data were transformed to make their distributions more normal in shape. Analyses of transformed scores produced the same effects as those of untransformed scores (see Appendix for details and source tables).

Interacting groups took an average of 583.5 seconds to reach a solution. The best co-acting individuals took an average of 435.3 seconds to reach a solution. Again, only groups that actually got a given rule right were included in this analysis. The difference was statistically significant ($p < .025$).

So, given that interacting groups and the best co-acting individual do not differ in performance, the co-acting individuals are more efficient problem-solvers: they take less time.

Standard repeated-measures analyses of time and card number were also performed--to see if time and cards to completion changed across the four rules. Those groups who did not solve a particular rule were assigned the time and card number at which they gave up. Because all groups had to play 60 cards, those groups who played 60 and could not make a guess were assigned a card-number of 61: it would have taken at least 61 cards for them to come up with a guess.

The problem with doing the standard repeated-measures analysis of all these scores is that groups who solved a given rule almost invariably took less time and cards than those groups that kept wrestling with the problem until close to the time-limit. Even though the difference was not significant, less co-acting individuals--particularly in the co-acting confirmatory cell--solved rules than interacting groups. A difference in time-to-completion, therefore, reflects a difference in solutions as well as a difference in times, because groups that failed to solve tended to take much longer.

But the analysis is valuable as a supplement to the time and card-to-solution analyses. Also, it provides another check on task difficulty: if the tasks are increasingly difficulty, we would expect them to take more cards and time to solve as we progress from one to four.

There is a significant task effect on cards-to-completion ($F(3,108) = 9.7$, $p < .0001$) and on time-to-completion ($F(3,108) = 19$, $p < .0001$). (See Appendix for source tables.) A look at the means shows that the rules did generally increase in difficulty:

	Rule 1	Rule 2	Rule 3	Rule 4
Cards-to-completion	31	35	41	49
Time-to-completion (in seconds)	433	537	689	670

The last two rules do not differ in the amount of time it takes to solve them, but otherwise, time and number of cards increase in linear fashion across the four rules. The rules do appear to increase in difficulty, although fatigue effects may play a role in decreased performance, especially on the last rule.

The time analysis also showed the same main-effect that was revealed by the time-to-solution analysis. Interacting groups do take longer than the best co-acting individuals, even when time-to-completion is used as the criterion ($F(1,36) = 4.26$, $p < .05$).

Interacting Groups Vs. the Best Co-Actor Across All Four Rules

Steiner (1972) predicted that interacting groups would perform better than the best of an equal number of individuals on a divisible task. If the four Eleusis rules are treated as parts of a single, divisible task, interacting groups should be more successful than the best co-actor, when 'best' is defined as the individual in each co-acting group who is most successful across the four rules.

For the present comparison, a single co-actor was selected from each group: the one who got the most rules right, using the least amount of time and cards (if two or more individuals solved the same number of rules). One other criterion was used. In one or two cases, one individual solved the first or second rule and another solved the third or fourth. Since rules three and four are clearly more difficult than one and two, in cases where two or more co-actors were tied in number of solutions within the group the best co-actor was the one who solved the higher rules.

The difference between the performance of the interacting groups and the best co-actor across all four rules is marginally significant ($F(1,36) = 2.93$, $p < .095$). (See Appendix for source table.)

Comparing interacting groups and the best co-actor across four tasks on the time- and cards-to-completion dependent measures revealed no significant differences (see Appendix for source tables). The best co-actor across the set of problems requires the same amount of time and information to see each solution as an interacting group.

Do Interacting Groups Solve a Higher Percentage of Problems Than Co-Actors?

This comparison corresponds to the one used by Marjorie Shaw in her classic (1932) study of differences between group and individual problem-solving. Here is a breakdown of percentages of correct solutions, task-by-task:

Rules	Interacting Groups Who Solved		Co-Acting Solvers	
	Total Interacting Groups		#Co-Acting Individuals	
1	18/20	90%	32/80	40%
2	17/20	85%	35/80	44%
3	16/20	80%	24/80	30%
4	12/20	60%	13/80	16%

There is no statistical test that can be used to determine whether these differences arose by chance--but this kind of difference has been found consistently in studies over the years (Davis, 1969; Steiner, 1972). It is a stable, reliable finding. In this study, interacting groups were successful 79% of the time; co-acting individuals were successful only 33% of the time.

So, interacting groups solved more Eleusis rules than co-acting individuals. Interacting groups also solved more rules than the best of an equal number of co-actors when 'best' means most successful *across all four rules*. But interacting groups were no more successful than the best of an equal number of co-actors when 'best' means most successful *on each rule*.

Mistakes

The number of cards each group got wrong was analyzed using a repeated-measures ANOVA. In this case, the dependent measure was number of cards wrong out of the total of sixty each group played on each rule.

Source	SS	df	F
C	6	1	.2
S	158	1	6.3* p<.02
CS	71.5	1	2.9
Er	900.1	36	
T	2907.8	3	46.2* p<.0001
CT	128.2	3	2
ST	197.5	3	3.1* p<.05
CST	10.7	3	.2
Er	63.6	108	

There is a significant main-effect for strategy. Disconfirmatory groups made an average of about twenty mistakes per rule, whereas confirmatory groups made only eighteen. The difference, although statistically significant, is quite small. There is also a significant strategy-by-task interaction, reflecting the fact that confirmatory groups made more mistakes on rule 1, but disconfirmatory groups made more mistakes on the other three rules.

	Rule 1	Rule 2	Rule 3	Rule 4
Confirmatory	26	15.2	19.6	12.6
Disconfirmatory	24.3	17.6	23.4	16
Overall	25.1	16.4	21.5	14.3

The scores in the table are all means. Note the large differences in number of cards wrong across the four rules: one and three produced many more mistakes than two and four. Rule four, in particular, was designed to make it easy to get cards right: subjects can alternate colors, or odds and evens, or play cards in order and never get a card wrong. Disconfirmatory groups got many more cards wrong on rule four than confirmatory groups; the difference on this rule alone was highly significant ($F(1,36) = 6.27$, $p < .017$). But there was no accompanying difference in performance.

When a regression is performed relating mistakes to correct solutions on rule four, r^2 equals .08. This means that mistakes are very poor predictors of success on rule 4. Across all four rules, mistakes are even worse predictors: $r^2 = .03$. Getting cards wrong apparently does not make it any easier for groups to arrive at the correct solution.

Other Factors That Affect Group Performance

The questionnaires subjects filled-out at the end of the experiment were analyzed to see if any of their answers predicted their performance. Two dependent variables were used in these regressions.

One was number of correct solutions. But this variable was computed in a way that reflected the relative difficulties of each rule. Rules one and two were assigned a value of one point apiece, rule three was assigned a value of two points and rule four was worth three points. The values reflect the fact that approximately equal numbers of groups solved rules one and two, many less solved rule three and still less solved rule four (see page 42). Why use points at all? Because a group that solved only rule four and a group that solved only rule one are not equal in performance: the group that solved rule four is clearly superior, as that rule is much more difficult.

When ANOVAs are done using points as the dependent measure instead of the repeated-measures approach to analyzing performance, there is still no difference in performance across the strategy and communication independent variables. (See Appendix for source tables.)

The second dependent-measure used in the regression analyses was group rank. All groups were ranked on two factors--the quality of their solutions, based on success points, and how many cards they used. The highest group, for example, took 61 cards to score seven points (get all four rules right).

This particular group was interacting, as were all of the top five groups. A Mann-Whitney U test indicated that the difference between the ranks achieved by interacting and co-acting groups was not significant, $U = 224$, $p < .12$.

The top group also included three seniors, each of whom had taken a large number of science and math courses. The worst group (co-acting and confirmatory) was composed entirely of freshmen who had almost no science and math background. They failed to get a single rule right.

Perhaps class rank and number of mathematics and science courses were related to success. These factors were used as lone predictors of success and rank, or combined to see if several variables taken together made better predictions than one variable alone.

There were four individual scores for class rank, science courses, etc. within each group. These scores were summed to produce a single group predictor score. Class rank, for example, was scored on a one through four scale, where one equals freshman and four equals senior. A group of four seniors would have a total class rank of sixteen. Number of mathematics and science courses were also summed within each group. Sums were used instead of individual scores because group members' performances were not independent: how one group member performed affected how all the others performed. In the interacting groups, all members made a single guess, together. In the co-acting groups, the best individual on each rule was used. But even his or her success was affected by the cards other group members played. A separate set of regressions were done using each co-acting individuals' rank and class scores to predict his or her success, and the predictive power was virtually nil (see Appendix for details).

A table of r^2s is presented below. Each r^2 represents the percentage of variance accounted for by each variable, or combination of variables. F-ratios were computed, comparing the amount of predicted variance to the amount of error variance. These F-values are not included in the table below, but p-values denoting their levels of significance are placed next to the appropriate r^2s.

Source	r^2 Success	r^2 Rank	
# Science Classes	.04	.09	
# Mathematics Classes	.07	.1*	$p<.05$
Class Rank	.08	.16*	$p<.01$
Class, Science and Math	.09	.17	
Science and Math	.08	.13	
Class and Science	.09	.16*	$p<.05$
Class and Math	.09	.16*	

No factor was very good at predicting group performance, in terms of number and quality of solutions. But class rank and number of mathematics classes, singly and together, managed to account for at least ten per cent of the variance on the rank dependent measure. It is true that seniors who have good science and mathematics backgrounds are likely to do better on the task than freshmen with little science background--but that prediction is very uncertain. What little relationship there is depends heavily on the fact that the first group had three senior science students in it and the fortieth group had four freshmen with no science background. Even in that top group, the one freshman made important contributions: the three seniors would not have done as well without her. It is clear that class rank, science and math background are only minor factors in achieving correct solutions to the four rules.

The questionnaire also asked subjects to rate their previous card-playing experience on a scale from one to five, where one equals almost no previous experience with card games and five equals extensive experience. Using card-playing experience to predict success produced an r^2 equal to .003. Card-playing experience was no help on the four rules.

Subjects were also asked to rate, on a scale from one (strongly disagree) to five (strongly agree) whether they agreed with the following statement: "If I am going to play cards I would rather play an easy game than a difficult thought game." Responses to this question were used to predict correct solutions; the resulting r^2 equaled .1. Those subjects that prefer difficult games do slightly, but not significantly, better on the four rules than subjects that prefer easier games.

When class rank and the difficult/easy question are combined as predictors of success, $r^2 = .17$. But even this r^2 is not significant even though it is higher than some of the r^2s on the previous page. The associated degrees-of-freedom are smaller--because the difficult/easy question only appeared on a later version of the questionnaire: about thirty of the forty groups saw it. The lower number of subjects meant that degrees-of-freedom were lower and the error-term was larger.

The questionnaires did not discover factors that accounted for a large part of the variance on the four Eleusis rules. Of the independent variables, only differences between the four rules seemed consistently related to number of correct solutions. In the next section, the implications of these results will be discussed.

IV. Discussion

The results of the present study replicate previous findings in the literature, using a novel task and design. Mynatt, Doherty and Tweney (1977) found that instructing subjects to use either a confirmatory or a disconfirmatory strategy did not affect their performance. Similar training also had no effect in the present study.

Marjorie Shaw's (1932) observation that interacting groups solve a higher percentage of problems than individuals working separately was replicated in the present study, with co-acting individuals playing the role of individuals working separately.

Steiner's (1972) argument that interacting groups should do better on a divisible task than the best of an equal number of individuals working separately was also supported. If interacting groups are compared with the best co-actor in each group across all four rules--with the rules constituting four phases of a single task--then interacting groups do perform slightly (but not significantly) better. In several interacting groups, one member solved rule three and communicated his or her solution to the others, making it possible for someone else to see the solution to rule four. Rule four is very hard to get if one has not seen the odd-even pattern on rule three.

The interacting/co-acting difference on this comparison might have been greater if interacting group members brought non-redundant resources to the task (Steiner, 1972). If each group member has special skills suited to a different aspect of a divisible task, then the group should perform much better than an equal number of individuals. Not even an Einstein could have developed the atomic bomb working alone, because the task required a coordinated effort on the part of chemists, physicists, engineers, politicians, etc.

Results also replicate the classic findings of Marquart (1955) and others (Steiner, 1972) that interacting groups perform no better and take more time on a single, non-divisible task than the best of an equivalent number of individuals working separately. According to the Lorge-Solomon model, on a Eureka-type task, the performance of interacting groups should equal $1 - Q^n$, where Q is the probability that one individual selected at random will not be able to solve the problem and n equals the size of the group. In the present study, we can estimate Q by taking the probability that a single co-acting individual will solve a given rule.

On rule four, for example, thirteen out of eighty co-acting individuals were successful, so Q equals .84. Then the probability that an interacting group will solve rule four equals $1 - (.84)^4$, or fifty per cent. Half the interacting groups, according to the Lorge-Solomon model, should have solved rule four. Actually, 60% did. The discrepancy is smaller on the other rules. On every rule, the Lorge-Solomon model predicts performance accurately.

The interacting/co-acting comparison bears an almost perfect resembance to the interacting/concocted comparison. Interacting groups performed exactly as we would have expected from the literature, and co-acting individuals performed in a manner similar to individuals working separately.

Co-acting groups do share some information. If one member plays cards sufficient to reveal the solution to one rule, it is quite possible that others will pick up on it. Still, in twenty-six out of the sixty solutions achieved by co-acting groups were the work of one person alone: no one else in the same group saw them.

Since there was no control group in which individuals worked on Eleusis alone, we cannot be certain that the best co-acting individuals would not have performed better than the best of a concocted group--but it is not likely, considering that the interacting/co-acting difference mirrors the interacting/concocting difference.

Why Do Interacting Groups Perform So Poorly?

Interacting groups clearly have an advantage over co-acting groups on tasks that require a division of labor and in situations where group members have non-redundant resources. But if an interacting group is poorly organized, members will not be able to take advantage of one another's resources and the group's product, even on a divisible task, will be no better than that of a co-acting group. This phenomenon is referred to as 'process loss' in the group problem-solving literature (Hoffman, 1979). The process by which group members try to arrive at a decision is often so inefficient that the group performs at a level far below its potential. That is why the Lorge-Solomon model tends to over-predict group performance on a Eureka-type task.

Steiner lists some reasons why groups might perform more poorly than expected on a Eureka task: "(1) The group will fail if none of its members possesses the resources demanded by the task. (2) The group will fail, or will function at a reduced level of effectiveness, if its processes are not in accord with task prescriptions. This will be the case if (a) the member(s) with the necessary resources does not use them to perform the unitary task; or (b) members with the necessary resources use them appropriately, but other members do not accept their contributions as the group's product (i.e., successful members are not accorded total weight)." (Steiner, 1972, p. 24). An example of a situation where process-loss occurs is a study by Torrance (1954). He asked B-26 bomber crews to solve a simple problem. The pilots, who were the commanders of the crews, were most successful at getting their opinions accepted, even when they were wrong. The leader or the majority in a group can ignore the opinions of those who are right.

It was hard for groups in the present study to organize, because most members began the sessions as strangers. One of the best interacting groups included two friends, one of whom became the group leader and the other of whom adopted a secretarial role, writing out the group's guesses. This group's organization was aided by the fact that the two friends slipped comfortably into their roles, the one dominant, the other submissive. In the group that did the best, three members were friends, and worked together smoothly.

But many groups in which the members were strangers did well and some in which there were friends did poorly. If the right atmosphere prevails and group members learn to work together, even strangers will be able to accomplish things they could not do separately. Perhaps the best example of this phenomenon is a group which included two foreign students who spoke little English. The two English-speaking students concentrated so much time and energy on explaining the rules to the two foreign students that they were forced to study their ideas carefully. The foreign students reciprocated the others' concentration. The result was a group in which everyone worked hard and made important contributions. This group missed the first rule, but got the other three. Had these individuals worked as a co-acting group, the foreign students would not have been able to understand the rules at all. Unfortunately, due to a procedural error by the experimenter, the data from this group had to be dropped from that analysis.

The five best groups, including three that got all four rules using less than a hundred cards, were from the interacting condition. A study of tapes and notes on these groups reveals only one thing they had in common: in all five, more than one group member contributed significantly to solving the rules.

In two of the worst five interacting groups, one member dominated discussion and lead the group on a false track: persistently looking for a color pattern on rule three in one case. More than one person tried to contribute in the other three 'worst groups,' but the group still fell into a 'set,' focusing on an irrelevant dimension. One group, for example, looked at differences between cards in terms of how many cards one had to 'skip.' A numerical difference of five is a 'skip' of four cards. Keeping track of the number of cards skipped proved too difficult: the group missed both rules three and four. If one member had seen that odd differences were the key to rule three, the group's performance would have been much better.

So, in a successful group, members have to criticize each other's approaches and look for alternatives. Otherwise, the group may fall into a harmful set and focus persistently on an irrelevant aspect of the cards.

Disconfirmatory instructions were an attempt to get group members to adopt a critical attitude. But the instructions failed: the best interacting groups were evenly divided between disconfirmatory and confirmatory conditions. Instructions in the present study focused on telling subjects to get cards wrong. Future studies should involve techniques for getting interacting group members to criticize each others' ideas and develop alternative hypotheses, in addition to making mistakes.

Group process can be beneficial, but it can also hurt performance. A set is much harder to form in a co-acting group; if one member focuses on an irrelevant dimension, another member--pursuing a different strategy--will often play a card that disconfirms the first member's idea. Not sharing information can have advantages as well as disadvantages. If one co-acting individual starts off on an irrelevant tangent, he or she doesn't take the whole group with him or her.

For example, in one interacting group, subjects deliberately played the same sequence of cards twice on rule three, to see if the rule required those cards to be played in exactly that order. The sequence was 4, K, 6, 7, 8, 7, 4 (the subjects were ignoring color) and it was played starting at card seven and later, starting at card twenty-nine. The rule in this case is odd-even, so clearly the sequence would work every time. All that was needed to break this sequence was a different pattern of odds and evens. In a co-acting group, there is a much greater chance that such a sequence would be broken immediately: if one member tried to play it, another member pursuing a different idea would disconfirm the first member's.

Why Disconfirmatory Groups Did No Better Than Confirmatory

Co-acting groups were less likely to fall into a harmful set because members would naturally tend to disconfirm one anothers' ideas. But co-acting groups also could not pursue a consistently disconfirmatory strategy. It was the interacting groups who had the greatest opportunity to take advantage of falsification. But the "Good Strategies for Guessing Patterns" had no effect on interacting groups' performance.

A rough inspection of groups' actual strategies, using tapes and notes, indicates that most groups followed a combined strategy--sometimes deliberately getting cards wrong, other times deliberately getting them right. Disconfirmatory groups in general made more deliberate mistakes (see Results, p. 49) indicating that most did try to follow their "Suggestions." But the difference in number of mistakes between confirmatory and disconfirmatory is not that great; if groups really followed their strategies, it should have been greater. Also, although most subjects indicated the "Good Strategies for Guessing Patterns" were helpful, the part many remembered best was the suggestion to "guess early, guess often."

Why did the suggestions have so little effect? On the questionnaire, only eighteen per cent of the subjects told to disconfirm mentioned specifically that getting cards wrong was helpful and only two per cent of the confirmatory subjects mentioned that getting cards right was helpful. The confirmatory instructions particularly made little impression on subjects. Part of the answer lies in the size of the individual players' card hands. Each member of a group was given thirteen cards from a shuffled deck at the beginning of each rule. If a group started playing a long string of cards that differed by one, sooner or later one member would not have the appropriate card--especially since group members were only given extra cards when they made mistakes. So that individual would have to play a card at random. The card might disconfirm the idea the group was working on. For example, many groups started out playing a long string of ascending cards on rule one, e.g., A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q. If the next player did not have a King, he or she might play a Jack, which would lead the group towards the real rule: adjacent cards must differ by one, regardless of order. A number of groups would have missed rule one were it not for a random play of this nature. The same kind of thing happened on other rules as well.

In fact, if a random card disconfirmed an incorrect guess on rule one, subjects were more likely to ignore confirmatory suggestions and try to disconfirm guesses on later rules. One of the advantages of using a four-task sequence is that subjects could learn better strategies for solving rules. By rule three or four, most groups had developed their own way--good or bad--of tackling Eleusis problems and were ignoring the oft-repeated suggestions.

The fact that each group had to play sixty cards on each rule exacerbated this effect. The probability that somewhere in the sixty cards someone would play a random card that revealed the rule was very large, even in the confirmatory groups. Often groups would fail to get the rule even when they had full information, but the playing of a random disconfirmatory card increased the possibility that a group would see the correct solution and that the group would deliberately play disconfirmatory cards on later rules.

So, the lack of a difference between strategy conditions can be partially explained by the fact that subjects had limited choices, in terms of what card to play, and they were forced to continue playing even if they made an early guess. 'Serendipity' is a term used in science to refer to a chance discovery. A classic example is the case of the mold that came in through the window and killed bacteria on one of Alexander Fleming's cultures. The mold produced Penicillin. Fleming was intelligent enough to make the most of his discovery. Similarly, many groups in the present study obtained serendipitous information from a card played at random. It was quite common to see a group keep on playing cards that differed by one on rule three until someone was forced to play something out of sequence, like a Queen after a three. The experimenter's "That's correct" was invariably followed by exclamations of surprise. Some groups were able to take advantage of this kind of information; others were not.

Disconfirmatory groups still should have done somewhat better than confirmatory. The "Suggestions" for disconfirmatory groups stressed looking at previous mistakes, as well as at right answers. In an alternating color rule, strings of mistakes will all be of the same color: red cards under a red card, black cards under a black card. In an odd-even rule, there will be long strings of odd mistakes under some odd cards and even mistakes under some even cards. Apparently, disconfirmatory groups did not make any better use of this information than confirmatory.

Disconfirmatory groups should also have falsified wrong ideas more rapidly than confirmatory; they should have required less time and cards to see the correct solution. A group that stumbles on disconfirmatory information by chance should discover it later than a group that is searching for it systematically.

In summary, the hypothesis that interacting-disconfirmatory groups would do better than interacting-confirmatory was not borne out by the evidence. But it still may be true. There are some procedural problems in the present study that increase the likelihood of serendipitous disconfirmation. What would happen if those problems were eliminated?

A Follow-Up Study

The author and several colleagues are currently conducting a study designed to follow-up this dissertation and discover if, under a different set of circumstances, there will be a difference between interacting-disconfirmatory and interacting-confirmatory groups. The procedures were the same as the ones used in the study described earlier except that:

a. Only interacting groups were run.

b. There was only one between-groups independent variable--strategy--with three conditions: confirmatory, disconfirmatory and a combined strategy, in which subjects were urged to get cards right until they had a guess, then test that guess by getting cards wrong.

c. Each subject was given a full deck of cards to play from.

Every time he or she played a card, it was immediately replaced. This meant that each subject could always play whatever card he or she wanted.

d. Groups were allowed to stop working on a rule whenever they were sure they were right.

Procedures c and d insured that confirmatory groups potentially could play a whole string of right answers without serendipitous disconfirmation, then quit and go on to the next rule without ever getting full, correct information. On rule three, for example, a confirmatory group could play thirty cards that went up and down by ones and stop, convinced that they had the rule--when the rule was actually odd and even cards must alternate.

The time and card number data have not been completely analyzed, but there is a strong difference between confirmatory and disconfirmatory strategies on number of correct solutions. Out of eight confirmatory groups run, one solved all four rules, one solved three, one solved one and the rest did not solve any rules. Furthermore, the groups that did the worst were the ones that followed their strategy the most closely, making almost no mistakes.

Out of eight disconfirmatory groups, three solved two rules, three solved three rules and two solved all four. The combined strategy group fell in between the confirmatory and disconfirmatory: one group solved all four, two groups solved three, three groups solved two and two solved one. A three-way ANOVA on correct solutions showed that the difference between these groups' performance on the strategy manipulation was significant, F(2,21) = 5.04, p < .02. If all groups had strictly followed their instructions, this difference would be even greater. Confirmatory groups tended to make more mistakes than they should have, and disconfirmatory groups got too many cards right.

It appears that interacting-disconfirmatory groups will do better than interacting-confirmatory when members are given maximum freedom to conduct their own experiments and stop whenever they feel they have the right answer. Confirmatory groups often fell into the trap of playing a short string of right answers and then quitting, thinking they had found the rule.

Extending These Results to Science

The dangers of generalizing from a sample of college students solving Eleusis rules to the behavior of scientists in the real-world were mentioned in the Introduction. Obviously, college students are not scientists and Eleusis is no more than a gross oversimplification of the kinds of problems most scientists actually face. Keeping these caveats in mind, let us speculate cautiously. What do these results suggest about how intragroup communication and the use of falsificatory strategies affect the scientific process?

1. On a Eureka-type task, the best of a number of scientists working separately and in limited communication with each other will do as well as scientists who work together in a face-to-face group. If the object is to solve a series of

related Eureka-type problems, an interacting group may have a slight advantage if individual resources are redundant (i.e., all group members have similar training and skills) and a large advantage if each individual has a special skill or perspective to contribute to one or more aspects of the problem. Finally, the *average* scientist working separately will not do as well as a group of scientists working together.

2. A disconfirmatory strategy will lead to superior performance in situations where scientists have the freedom to design whatever experiments they want. When laboratory subjects were given only a limited range of card-choices and were forced to play sixty cards, a disconfirmatory strategy had no effect. But when subjects could play any card they wanted, and stop any time, a disconfirmatory strategy was much more effective than a confirmatory one.

Similarly, when scientists are limited to certain tests and are forced--by a grant, perhaps--to gather data for an extended period of time, serendipitous disconfirmation of the researchers' hypotheses may occur. The limitations on kinds of tests may be imposed by lack of equipment, or by ethical considerations (certain kinds of research cannot be done on animals and human beings), or even by the nature of the universe--a geologist cannot go back in time to study the Earth's surface as it appeared millions of years ago. In a card-game, we can give subjects freedom to construct any test of a rule they want, but in the real world, scientists are forced to operate within certain constraints.

Still, experimenters are quite adept at arranging things so their pet hypotheses will be confirmed. A classic demonstration of the effect of rewards on behavior was done in the 1940s. Cats were put in a cramped box with a pole in the middle. When the cat rubbed the pole, the box opened. The experimenters concluded that the cats learned to rub because it brought about an immediate reward--release from the box. Thirty years later, two researchers designed a disconfirmatory study. They put cats in a similar box and gave them no reward. The cats rubbed the pole repeatedly, demonstrating that rewards had been unnecessary in the original experiment (Garcia, 1980). Subjects solving Eleusis rules and scientists trying to solve the 'rules of nature' can both profit by adopting a disconfirmatory strategy.

Future Research

To find out whether the ideas derived from the study of subjects in the laboratory really generalize to the behavior of scientists, it will be necessary to study scientists themselves. One way would be to bring scientists into the laboratory and have them work in interacting and co-acting teams on Eleusis rules. This kind of study should be supplemented by work with research teams in the real world, like Mitroff's (1974) investigation of the Apollo scientists. Under what circumstances is it better for scientists to wrestle with a problem together, and when are they better off working alone? Can scientific groups pursue a disconfirmatory strategy better than individuals?

More realistic laboratory simulations of science could also be done using college students. For example, one could take a group of junior and senior science majors and put them in a course designed to model the growth and development of a science.

The students would be asked to work--together and separately--on discovering the rules that govern a 'universe' created for the purposes of the experiment. The best way to design such a 'universe' would be to use a computer. Students could work at several different terminals, trying different experiments that would reveal the 'laws' governing the universe.

The experimenter could construct laws that dictated the behavior of particles in the imaginary universe--how they moved, how they interacted, how they combined to form larger bodies. This information would have to be discovered by the students, through clever experimentation. The simulation could be arranged so that it would be easy to form a simple theory of how the universe worked--a theory that would not hold up to rigorous, disconfirmatory testing.

Some of the contingencies that affect the progress of science in the real world could be modeled. Students' access to the computer terminals could be determined by fake 'funding agencies' that would reward only certain kinds of research--say, those experimenters whose predictions were confirmed. Students who did riskier projects and/or disconfirmatory studies would have difficulty getting access to terminals.

Academic journals could also be simulated; students could be told to write up the results of their experiments, and acceptable articles could be reproduced and distributed. The students with the most publications would also be permitted more time on the computer.

The list of possible permutations is endless. Simulations are far more realistic than short, little experiments. Simulations are also far more expensive and time-consuming. But the potential rewards are tremendous. One could actually model the growth of a science under circumstances where key factors in scientific progress could be manipulated to assess their effects. This kind of work could even involve actual scientists as consultants and participants.

Simulation is not the only way to study the scientific process, but it is a method whose possibilities have not been fully realized. One possibility is to give subjects monetary rewards for either getting cards right (confirmatory strategy) or getting them wrong (disconfirmatory strategy). Even in my most recent study, groups showed a tendency to disregard their strategy instructions. Perhaps a monetary contingency would convince groups to adhere more closely to their strategies, and give a clearner test of the differences between confirmatory and disconfirmatory strategies.

Another possible way of obtaining the same information would be to ask groups to concentrate solely on either getting cards right or getting them wrong. After each group had played a certain number of right or wrong cards in a row (depending on which strategy it had been asked to adopt) the experimenter could ask the group to guess the rule. Groups that could not follow their strategy would be dropped from the analysis. Note that the focus here is on playing long strings of right or wrong cards, not on guessing the rule. This design would permit an assessment of how well subjects could create and make use of pure confirmatory or disconfirmatory information. While several groups in my most recent study played a pure confirmatory strategy, no group has ever played a pure disconfirmatory strategy.

Other ideas for future research include a questionnaire that assesses how well scientists employ modus tolens on Eleusis rules and a computer program that plays Eleusis, which a colleague of mine is working on. The computer's performance using confirmatory and disconfirmatory strategies could be compared with that of human subjects.

The experiment outlined in these pages has already begun to fulfill its primary purpose: to suggest new ways of simulating how science works in the laboratory. Another possible offshoot of the research presented here is techniques to improve group problem-solving in general. A good strategy for a group might be to generate ideas in a brainstorming session, then attempt to disconfirm each idea, using individual group member's unique resources and backgrounds to come up with falsificatory evidence. The efficacy of this and other group problem-solving techniques can be investigated in further experimental simulations, coupled with observations of groups outside the laboratory.

APPENDIX

APPENDIX

ADDENDA TO THE RESULTS SECTION

Interacting Groups vs. the Best Co-Actor on Each Rule

ANOVA on number of correct solutions, Rule 4: (See page 43 for discussion)

Source	SS	df	F
C	.1	1	.4
S	.4	1	1.6
CS	.4	1	1.6
Er	9.0	36	

ANOVA on the number of cards required to reach a solution: (See page 43 for discussion)

Source	SS	df	F
C	15.7	1	.05
S	93.8	1	.33
CS	215.3	1	.75
Er	33871.0	118	

ANOVA on cards-to-completion: (See page 46 for discussion)

Source	SS	df	F
C	65	1	.1
S	140	1	.2
CS	722	1	1.3
Er	19998	36	

Source	SS	df	F
T	7532	3	9.7* p<.009
CT	247	3	.3
ST	494	3	.6
CST	158	3	.2
Er	27891	108	

To make the card-number distribution approximate normality, each score was transformed by taking its logarithm to the base ten. The following source tables resulted:

Cards-to-solution:

Source	SS	df	F
C	.0004	1	.006
S	.04	1	.56
CS	.13	1	.98
Er	7.69	119	

Cards-to-completion:

Source	SS	df	F
C	.007	1	.08
S	.052	1	.57
CS	.253	1	2.75
Er	3.311	36	
T	1.6	3	10.03* p<.001
CT	.05	3	.32
ST	.07	3	.45
CST	.1	3	.61
Er	5.77	108	

The pattern of results is no different after transformation.

Time-to-completion: (See page 44 for discussion)

Source	SS	df	F
C	1158891	1	4.26* p<.05
S	1080	1	.004
CS	298166	1	1.1
Er	9782093	36	
T	7844208	3	19.04* p<.001
CT	282302	3	.68
ST	443671	3	1.077
CST	155527	3	.38
Er	14830259	108	

A square-root transformation was used to make the distinction of times approximate normality. The following source tables resulted:

Time-to-solution:

Source	SS	df	F
C	275	1	4.64* p<.05
S	1.3	1	.022
CS	2.6	1	.043
Er	7075.4	119	

Time-to-completion:

Source	SS	df	F
C	373.7	1	3.49^{+}
S	.4	1	.004

Source	SS	df	F
CS	109.	1	1.02
Er	3854.7	36	
T	2911.	3	18.33* p<.001
CT	30.5	3	.19
ST	138.4	3	.87
CST	14.3	3	.09
Er	5716	108	

The (+) denotes a marginally significant communication main-effect (p < .07). So, the transformation slightly reduces the effect of the communication variable on time, but does not affect the overall pattern of results.

Interacting Groups vs. the Best Co-Actor Across All Four Rules

ANOVA on number of correct solutions: (See page 47 for discussion)

Source	SS	df	F
C	.75	1	2.93^+
S	.05	1	.22
CS	.51	1	1.96
Er	9.28	36	

The (+) denotes a marginally-significant communication main-effect (p < .095). The within-group part of the analysis showed the usual main-effect for task--later rules are more difficult to solve than earlier ones--and no interactions.

On page 47, the reader is referred to this appendix for source tables comparing interacting groups and the best co-actor across the four rules on time- and card-to-solution. Only the source tables for transformed data will be reported here, because the transformed time and card scores come closer to satisfying the assumptions of the analysis-of-variance. (Analyses were also performed on the raw scores; the F-ratio for the communication main-effect on the time dependent measure was slightly higher than the same F-ratio for square-root-time--but neither was significant at the .05 level.)

Square-root of time-to-solution:

Source	SS	df	F
C	168.13	1	2.79
S	1.89	1	.03
CS	.85	1	.01
Er	6740.93	112	

Logarithm to the base ten of cards-to-solution:

Source	SS	df	F
C	.008	1	.14
S	.001	1	.02
CS	.04	1	.65
Er	6.83	112	

Using points as a dependent-measure

On pages 50 and 51, a system of points that weights each rule according to its relative difficulty is described. When these points are used as the dependent-measure in a 2 X 2 ANOVA, the following source table results:

Source	SS	df	F
C	2.5	1	.59
S	1.6	1	.38
CS	8.1	1	1.8
Er	153.4	36	

None of these F-ratios comes close to significance. When the same ANOVA is performed using the point-totals for the best co-actor across the four rules, there are still no significant differences.

Regressions using individual co-actors' scores: (See page 52 for discussion)

Source	r^2
Class Rank	.02
Science Courses	.0007
Math Courses	.01

REFERENCE NOTES

Reference Notes

[1] Kern, L. H., Mirels, L. H., & Hinshaw, V. G. Scientists' understanding of propositional logic: An experimental investigation. Paper presented at the Meeting of the American Psychological Association, Montreal, September, 1980.

[2] Spencer, N. J., Hartnett, J., & Mahoney, J. Quality of reviewing by journal referees. Paper presented at the meeting of the Eastern Psychological Association, Hartford, April, 1980.

LIST OF REFERENCES

List of References

Agnew, N. M., & Pyke, S. W. The science game. Englewood Cliffs, N.J.: Prentice-Hall, 1969.

Davis, J. H. Group performance. Reading, MA: Addison-Wesley, 1969.

Einhorn, H. J., & Hogarth, R. M. Confidence in judgement: Persistence of the illusion of validity. Psychological Review, 1978, 85, 395-416.

Einstein, A., & Infeld, L. The evolution of physics. Forge Village, MA: Simon & Schuster, 1938.

Garcia, J. Tilting at the paper mills of academe. American Psychologist, 1981, 36, 109-116.

Gardner, M. On playing New Eleusis, the game that simulates the search for truth. Scientific American, 1977, 237(4), 18-25.

Gould, S. J. Ever since Darwin. New York: W. W. Norton, 1977.

Hoffman, R. L. Applying experimental research on group problem-solving to organizations. Journal of Applied Behavioral Science, 1979, 15, 353-391.

Holton, G. Thematic origins of scientific thought. Cambridge: Harvard University Press, 1973.

Judson, H. F. The eighth day of creation. New York: Simon & Schuster, 1979.

Kaufmann, W. J. Relativity and cosmology. New York: Harper & Row, 1973.

Kelley, H. H., & Thibaut, J. W. Group problem solving. In G. Lindzey & E. Aronson (Eds.), The handbook of social psychology (Vol. 4). London: Addison-Wesley, 1969.

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, CA: Brokks/Cole, 1968.

Koestler, A. The sleepwalkers. New York: Grosset & Dunlap, 1963.

Kuhn, T. S. The structure of scientific revolutions (2nd ed.). Chicago: University of Chicago Press, 1970.

Mahoney, M. J. *Scientist as subject*. Cambridge: Ballinger, 1976.

Mahoney, M. J., & Kimper, T. P. From ethics to logic: A survey of scientists. In M. J. Mahoney, *Scientist as subject*. Cambridge: Ballinger, 1976.

Marquart, D. I. Group problem-solving. *Journal of Social Psychology*, 1955, *41*, 103-113.

May, H. G., & Metzger, B. M. (Eds.). *The Oxford annotated Bible*. New York: Oxford University Press, 1965.

Mitroff, I. I. *The subjective side of science*. Amsterdam: Elsevier, 1974.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. Confirmation bias in a simulated research environment: An experimental study of scientific inference. In P. N. Johnson-Laird & P. C. Wason (Eds.), *Thinking: Readings in cognitive science*. Cambridge: Cambridge University Press, 1977.

Popper, K. R. *Conjectures and refutations*. New York: Basic Books, 1962.

Popper, K. R. *Unended quest*. La Salle, IL: Open Court, 1976.

Romesburg, H. C. Simulating scientific inquiry with the card game Eleusis. *Science Education*, 1979, *5*, 599-608.

Shaw, M. E. A comparison of individuals and small groups in the rational solution of complex problems. *American Journal of Psychology*, 1932, *44*, 491-504.

Sherwin, M. J. *A world destroyed*. New York: Vintage Books, 1977.

Skyrms, B. *Choice and chance: An introduction to inductive logic*. Belmont, CA: Dickenson, 1966.

Steiner, I. D. *Group process and productivity*. New York: Academic Press, 1972.

Torrance, E. P. Some consequences of power differences on decision making in permanent and temporary three-man groups. *Research Studies*, *State College of Washington*, 1954, *22*, 130-140.

Wason, P. C. "On the failure to eliminate hypotheses"...a second look. In P. N. Johnson-Laird & P. C. Wason (Eds.), *Thinking: Readings in cognitive science*. Cambridge: Cambridge University Press, 1977.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.