Scientific Induction: Individual versus Group Processes and Multiple Hypotheses

Eric G. Freedman
Department of Humanities
Michigan Technological University
Houghton, MI 49931-1295
freedman@mtus5.bitnet

Abstract

It has been suggested that groups can evaluate multiple hypotheses better than individuals. The present study employed Wason's (1960) 2-4-6 task to examine the effects of multiple hypotheses in scientific induction. Subjects worked either individually or in four-member interacting groups. Subjects were also instructed to test either a single or a pair of hypotheses. The results indicate that groups perform significantly better than individuals. When testing multiple hypotheses, groups were more likely to determine the target hypothesis than individuals. Interacting groups generated more positive tests that received negative feedback and received more disconfirmation than individuals. When multiple hypotheses were tested, interacting groups used greater amounts of diagnostic tests than individuals. Interacting groups appear to search their experiment space and evaluate the evidence received better than individuals.

Introduction

In science as well as in everyday induction, people have been shown to rely on a positive-test strategy, that is, they tend to generate tests intended to confirm their hypotheses (see, Klayman & Ha, 1987, for review). Although a negative-test strategy (i.e., disconfirmatory) has been assumed to facilitate induction, subjects are often unable to benefit from this strategy (Freedman, 1991a; Gorman & Gorman, 1984; Tweney et al., 1980). Farris and Revlin (1989) have suggested that subjects may not benefit from a negative-test strategy as a result of an inability to consider alternate hypotheses.

In previous research, subjects have typically worked individually to test a single hypothesis (Gorman & Gorman, 1984; Gorman, Gorman, Latta, & Cunningham, 1984; Hacker, Freedman, Gorman, & Isaacson, 1990; Wason, 1960). Platt (1964) has claimed that scientific induction is facilitated when several hypotheses are tested simultaneously and particular hypotheses are eliminated through experimentation. In fact, Klayman and Ha (1987) have suggested that the evaluation of multiple hypotheses remains an important area for further research. Unfortunately, studies, which encouraged subjects to consider multiple hypotheses, have produced mixed results. Whereas Tweney et al. (1980, Experiment 2) found that encouraging subjects to use multiple hypotheses reduced performance, Klahr and Dunbar (1988, Experiment 2) and Klayman and Ha (1989) found that asking subjects to consider alternative hypotheses improved performance. Yet, McDonald (1990) found that multiple hypotheses were effective only when the target hypothesis was a subset of subjects' initial hypothesis. Freedman (1991b) found that multiple hypotheses improved performance only when used in conjunction with a negative-test strategy. Moreover, Freedman (1991a, 1991b) and Klahr, Dunbar, and Fay (1990) found that subjects employing multiple hypotheses generated significantly fewer experiments than subjects testing a single hypothesis. Thus, multiple hypotheses may reflect a more efficient strategy than single hypotheses.

Whereas individuals have difficulty testing more than one hypothesis at a time, Gorman (1986) has hypothesized that groups "can keep track of several hypotheses at once" (p. 93). Laughlin and Futoran (1985) found that groups performed better than individuals. Gorman et al. (1984, Experiment 1) found that interacting groups determined the target hypothesis as well as the best members of non-interacting groups. Laughlin and Futoran found that groups did not form better hypotheses than individuals; however, groups did evaluate hypotheses better than individuals. Nevertheless, Freedman (1991a)

found that even though encouraging individual members of interacting groups to consider alternative hypotheses did not facilitate induction, successful groups who tested multiple hypotheses conducted significantly fewer experiments than successful groups who evaluated a single hypothesis at a time.

In short, whereas several studies seem to favor the testing of multiple hypotheses over single hypotheses, the overall results are inconclusive. It is therefore important to determine more precisely when multiple hypotheses facilitate scientific induction (Klayman & Ha, 1987). Multiple hypotheses are assumed to be effective when they permit a more extensive search of the hypothesis and experiment problem spaces (Klahr, Dunbar, & Fay, 1990). The present experiment used Wason's (1960) 2-4-6 task to investigate two dimensions simultaneously: single versus multiple hypotheses and four-member interacting groups versus individuals working alone. Multiple hypotheses have not been shown to facilitate induction, in part, because multiple-hypotheses studies have typically been run on individual subjects working alone. Therefore, this study examined whether groups can evaluate multiple hypotheses better than individuals. Another reason why multiple hypotheses may not routinely enhance induction is due to the fact that subjects have difficulty mentally representing more than one hypothesis at a time (Freedman, 1991a). In order to make alternate hypotheses more concrete, subjects were required to state a pair of hypotheses on each trial. Consistent with previous group problem-solving research, interacting groups are predicted to perform better than individuals. If interacting groups are better than individuals at evaluating alternative hypotheses, groups should be more likely to determine the target hypothesis when testing multiple hypotheses.

Method

Subjects. One hundred-twenty students enrolled in introductory psychology classes at Michigan Technological University participated in this study. All subjects received course credit for participation.

Procedure. Each group of three to five people was randomly assigned to either the individual-subjects condition or the four-member interacting group condition and to either the single-hypothesis or the multiple-hypotheses conditions. In the single-hypothesis (**SH**) condition, subjects proposed a single hypothesis and a number sequence to test it. In the multiple-hypothesis (**MH**) condition, subjects generated a pair of hypotheses and a number sequence to test them. They were read instructions very similar to those used in previous research (Gorman & Gorman, 1984). Subjects were told that the sequence, 2-4-6, is an instance of a target hypothesis and that they had to determine the target hypothesis by proposing other number sequences. The target hypothesis was—any three different numbers.

For each trial, subjects recorded their hypothesis or hypotheses, number sequence, whether their sequence conformed to their hypothesis and the experimenter's feedback. Subjects were given feedback regarding whether their number sequence was consistent with the target hypothesis but they were not told whether their hypotheses were correct. Prior to the main task, subjects were given a four-trial practice problem. Finally, subjects terminated the experiment when they believed they had determined the target hypothesis. However, a 25-minute time limit was imposed on the main task.

Results

A 2 X 2 (Number of Hypotheses X Number of Subjects) ANOVA was computed on each dependent variable. Differences between subjects who did and did not determine the target hypothesis were also analyzed.

Solutions. The proportion of subjects (based on 12 subjects/cell) who successfully discovered the target hypothesis is presented in Figure 1. Although neither the main effect of Number of Hypotheses or the two-way interaction reached significance, the main effect of the Number of Subjects indicated that interacting groups were more likely to determine the target hypothesis than individuals, $F(1,44) = 19.57$, $p < .0001$. When multiple hypotheses were evaluated, groups were more likely than individuals to discover the target hypothesis.

Experiments. Because subjects terminated the task when they believed they had discovered the target hypothesis, the total amount of number sequences proposed was measured. SH subjects conducted more experiments than MH subjects ($M = 16.13$ vs. 10.49), $F(1,44) = 15.37$, $p < .0005$. Subjects testing a single hypothesis determined the target hypothesis as often as subjects testing multiple hypotheses because SH subjects gathered more information than MH subjects. In addition, for successful subjects (i.e. subjects who determined the target hypothesis), MH subjects announced the target hypothesis earlier than SH

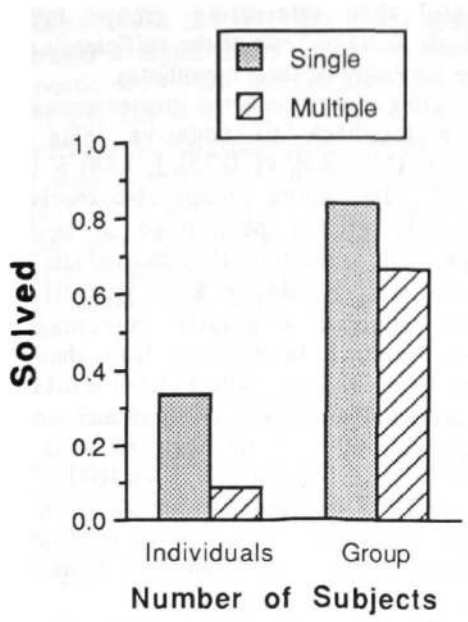

Figure 1
Probability of determining the target hypothesis

subjects (M = 9.56 vs. 12.57), F(1,19) = 4.70, p < .05. Thus, these results provide further support for the idea that a MH strategy appears to be more efficient than a SH strategy.

Hypotheses. Because subjects explicitly stated a hypothesis or a pair of hypotheses on each trial, the total number of different hypotheses proposed was measured. MH subjects generated significantly more unique hypotheses than SH subjects (M = 9.08 vs. 5.38), F(1,44) = 11.03, p < .002. This finding provides some preliminary support for the idea that testing multiple hypotheses led to an increased search of the hypothesis space. The average number of tests conducted for each hypothesis was also measured. SH subjects also conducted a greater number of tests per hypothesis than MH subjects (M = 3.93 vs. 1.20), F(1,44) = 23.34, p < .0001. Thus, subjects testing a single hypothesis maintain their hypotheses longer than those testing multiple hypotheses. It also suggests that subjects testing a single hypothesis rely more on a search of the experiment space than subjects testing multiple hypotheses.

Test Strategy. The number and percentage of sequences that conformed (i.e., positive tests) or did not conform (i.e., negative tests) to subjects' hypotheses was gathered. For the number and the percentage of positive and negative tests, none of the effects reached statistical significance. However, successful subjects proposed a significantly greater amount of negative tests than unsuccessful subjects (M = 8.96 vs. 6.28), F(1,44) = 8.90, p < .005. Apparently, successful subjects in the present study are able to benefit from the use of a negative-test strategy.

Experimenter's Feedback. Number sequences which were either consistent (i.e., positive feedback) or inconsistent (i.e., negative feedback) with the target hypothesis were recorded. SH subjects received greater amounts of positive feedback than MH subjects (M = 13.92 vs. 9.29), F(1,44) = 7.86, p < .01. SH subjects also received a greater percentage of positive feedback than MH subjects (M = .828 vs. .455), F(1,44) = 152.58, p < .0001. Individuals received a greater percentage of positive feedback than interacting groups (M = .695 vs. .587), F(1,44) = 12.80, p < .001. MH subjects also received significantly less negative feedback than SH subjects (M = 1.00 vs. 2.88), F(1,44) = 10.94, p < .0005. Interacting groups received more negative feedback than individuals (M = 3.00 vs. 0.875), F(1,44) = 14.05, p < .002. Furthermore, successful subjects received greater amounts of negative feedback than unsuccessful subjects (M = 3.91 vs. 0.12), F(1,44) = 42.17, p < .0001. This pattern of results suggests that subjects testing multiple hypotheses and interacting groups conduct a more extensive search of the experiment space.

Confirmation. The amount of confirmation was calculated by combining positive tests that received positive feedback and negative tests that received negative feedback. The amount of disconfirmation was calculated by combining positive tests that received negative feedback and negative tests that received positive feedback. Although no significant differences were observed in the amounts or percentages of confirmation received, a Success X Number of Subjects interaction indicated that successful interacting groups and unsuccessful individuals received a relatively greater percentage of confirmation than the other groups, F(1,44) = 5.59, p < .05. A Success X Number of Hypotheses interaction indicated that successful SH subjects and unsuccessful MH subjects received a relatively greater percentage of confirmation than the other groups, F(1,44) = 5.18, p < .05. Successful interacting groups may benefit from additional confirmation because of their superior hypothesis evaluation abilities.

A Number of Subjects X Number of Hypotheses interaction indicated that MH interacting groups received greater amounts of

disconfirmation than the other conditions (see Figure 2), $F(1,44) = 4.25$, $p < .05$. Subjects in the MH condition received significantly more disconfirmation than subjects in the SH condition ($M = 8.25$ vs. 5.50), $F(1,44) = 11.58$, $p < .002$. A Number of Subjects X Number of Hypotheses interaction indicated that MH interacting groups received a greater percentage of disconfirmation than the other conditions, $F(1,44) = 4.17$, $p < .05$. Thus, one reason why interacting groups may be able to evaluate multiple hypotheses better than individuals is that interacting groups receive more disconfirmation than individuals. Disconfirmation may help interacting groups to eliminate incorrect alternate hypotheses.

Individuals conducted a significantly greater percentage of positive tests resulting in positive feedback than interacting groups ($M = .586$ vs. .486), $F(1,44) = 5.29$, $p < .05$. Successful subjects received a smaller percentage of positive feedback to positive tests compared with unsuccessful subjects ($M = .449$ vs. .616), $F(1,44) = 11.88$, $p < .002$. A Success X Number of Subjects interaction indicated that unsuccessful individuals received a relatively higher percentage of this type of confirmation than the other subjects, $F(1,44) = 4.28$, $p < .05$. According to Klayman and Ha (1987), this type of test allows subjects to determine the sufficiency of their hypotheses. Thus, individuals may be less successful than interacting groups because individuals focus on tests of the sufficiency rather than the necessity of their hypotheses.

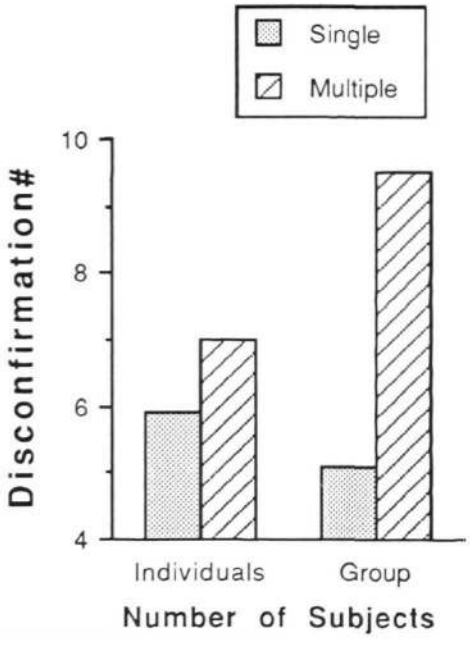

Figure 2
Mean amount of disconfirmation

Interacting groups received greater amounts of negative feedback to negative tests than individuals ($M = 2.50$ vs. 0.75), $F(1,44) = 10.35$, $p < .005$. Interacting groups also received a significantly greater percentage of negative feedback to negative tests than individuals ($M = .134$ vs. .043), $F(1,44) = 8.52$, $p < .01$. SH subjects received a greater percentage of negative feedback to negative tests than MH subjects ($M = .128$ vs. .049), $F(1,44) = 6.41$, $p < .02$. Successful subjects received greater amounts of negative feedback to negative tests than unsuccessful subjects ($M = 3.39$ vs. 0.00). Thus, not all types of confirmation are detrimental to scientific induction. This type of confirmation may allow subjects to determine the limits of the target hypothesis.

For positive tests that received negative feedback, interacting groups received significantly greater amounts of this type of information than individuals ($M = 1.38$ vs. 0.25), $F(1,44) = 8.92$, $p < .01$. Interacting groups also received a greater percentage of negative feedback to positive tests than individuals ($M = .064$ vs. .018), $F(1,44) = 6.88$, $p < .02$. Furthermore, successful subjects received greater amounts of negative feedback to positive tests than unsuccessful subjects ($M = 1.57$ vs. 0.12), $F(1,44) = 8.04$, $p < .01$. Successful subjects also received a significantly greater percentage of negative feedback to positive tests than unsuccessful subjects ($M = .075$ vs. .010), $F(1,44) = 7.33$, $p < .01$. As Hoenkamp (1989) has suggested, groups may be able to use positive tests more effectively than individuals because groups generate experiments that help them to decide between their hypotheses and the target hypothesis. For negative tests which received positive feedback, MH subjects generated more of this type of this type of test than SH subjects ($M = 7.25$ vs. 4.71), $F(1,44) = 13.20$, $p < .001$.

Diagnostic Tests. Because subjects can not know beforehand whether their hypotheses will be disconfirmed, the only sure way to disconfirm a hypothesis, when multiple hypotheses are tested, is to conduct a diagnostic test. In the MH condition, diagnostic tests were measured by counting each test that was an instance of one hypothesis and was not an instance of the other one. Clearly, subjects appreciated the importance of diagnostic tests. Diagnostic tests were employed on 62% of the trials. More

diagnostic tests were generated by interacting groups than by individuals (M = 7.92 vs. 4.50), $F(1,22) = 18.81$, $p < .01$. Interacting groups also generated a significantly greater percentage of diagnostic tests than individuals (M = .754 vs. .481), $F(1,22) = 10.47$, $p < .005$. Successful subjects produced greater amounts of diagnostic tests than unsuccessful subjects (M = 8.33 vs. 4.93), $F(1,22) = 5.79$, $p < .05$. Thus, another reason why groups may evaluate multiple hypotheses better than individuals is because groups conduct more diagnostic tests.

Discussion

Interacting groups were able to determine the target hypothesis more often than individuals. When subjects employed multiple hypotheses, groups performed better than individuals. The results of the present study provide further support for the view that individuals may have difficulty forming a mental representation of alternate hypotheses and therefore they are not able to benefit from the presence of multiple hypotheses (Freedman, 1991a). Consistent with my previous research (Freedman, 1991a, 1991b), testing multiple hypotheses does not increase the overall likelihood of determining the target hypothesis, but MH subjects generate significantly fewer number sequences than subjects in the SH condition. Additionally, successful MH subjects announce the target hypothesis sooner than successful SH subjects. Thus, testing multiple hypotheses enables discovery of the target hypothesis more efficiently than testing a single hypothesis. As Hoenkamp (1989) has suggested, rather than emphasizing the use of confirmatory or disconfirmatory strategies, the optimal strategy may be one which minimizes the number of experiments conducted.

Evaluating multiple hypotheses may be an efficient strategy during scientific induction because it promotes the elimination of incorrect hypotheses. This conclusion was supported by the finding that MH subjects proposed significantly fewer tests per hypothesis compared with SH subjects. MH subjects also received more disconfirmation and generated more negative tests that received positive feedback than SH subjects. Thus, the presence of alternate hypotheses appears to make the possibility of disconfirmation more salient because it forces subjects to consider the necessity as well as the sufficiency of their hypotheses. When testing a single hypothesis, subjects tend to focus on the sufficiency of their hypotheses as reflected in a relatively greater reliance on positive tests that receive positive feedback. Higher levels of negative feedback in the presence of multiple hypotheses suggests that multiple hypotheses result in a more extensive search of the experiment space. This conclusion is based on the fact that it is harder to obtain negative feedback when a general target hypothesis is employed. Interacting groups may be able to evaluate multiple hypotheses better than individuals because groups receive a greater amount and a higher percentage of disconfirmation. This disconfirmation allows groups testing multiple hypotheses to eliminate incorrect hypotheses. Thus, interacting groups do benefit from receiving disconfirmation. Furthermore, a greater reliance on a diagnostic strategy may facilitate interacting groups' utilization of multiple hypotheses because this type of test provides the best strategy to eliminate incorrect hypotheses.

The results of the present study indicate that interacting groups do not generate greater amounts of number sequences, positive tests, and negative tests. Thus, groups do not receive more information than individuals. Nor, do groups differ in their *intended* strategies (i.e., positive- versus negative-test strategies). In other words, interacting groups do not seek more confirmation or disconfirmation than individuals. Rather, interacting groups are superior in evaluating the information they receive. Nevertheless, it may be the case that groups generate more informative tests (Hoenkamp, 1989) than individuals. Indeed, as Hoenkamp has suggested, a formal analysis of the informativeness of particular tests may yield further insights into the evaluation of multiple hypotheses by individuals and interacting groups. Still, the fact that interacting groups received more negative feedback, less positive feedback, more negative feedback to negative tests, more negative feedback to positive tests, and a smaller percentage of positive feedback to positive tests than individuals indicates that the information interacting groups receive differs from the information individuals receive. Once more, with a broad target hypothesis, negative feedback to various types of tests requires that groups conduct a more extensive search of the experiment space. Interacting groups may conduct a more extensive search of the hypothesis space than individuals because group members typically propose several number sequences from which the most informative sequence is chosen. Often, the most informative number sequence is the one that diverges from previous tests. Furthermore, a greater reliance on tests which lead to disconfirmation may help interacting groups to

abandon incorrect alternate hypotheses.

The results of the present study suggest that the way in which subjects search through the experiment and hypothesis space is more important than whether subjects seek confirmation or disconfirmation of their hypotheses. When multiple hypotheses are evaluated, an extensive search of the hypothesis and experiment space may allow these subjects to discover the target hypothesis more expeditiously because they can determine the boundaries of the target hypotheses as well as eliminate incorrect alternative hypotheses. Still, the present study did not attempt to influence the types of hypotheses proposed or the experiments conducted. It is quite possible that encouraging subjects to propose maximally different hypotheses and to conduct diagnostic tests would facilitate scientific induction. While this study alone can not provide a comprehensive explanation of the conditions under which multiple hypotheses facilitate scientific induction, the present study does provide further evidence that when the cost of conducting experiments is high, use of multiple hypotheses may be preferable to the use of a single hypothesis.

References

Farris, H., and Revlin, R. 1989. Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. *Memory and Cognition* 17:221-232.

Freedman, E. G. 1991a. Role of multiple hypotheses during collective induction. Paper presented at the 61st Annual Meeting of the Eastern Psychological Association New York, NY.

Freedman, E. G. 1991b. Scientific induction: Multiple hypotheses and test strategy. Paper presented at the 32nd Annual Meeting of the Psychonomic Society, San Francisco, CA.

Gorman, M. E. 1986. How the possibility of error affects falsification on a task that models scientific problem solving. *British Journal of Psychology* 77:85-96.

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

Gorman, M. E., Gorman, M. E., Latta, M., and Cunningham, G. 1984. How disconfirmatory, confirmatory and combined strategies affect group problem solving. *British Journal of Psychology* 75:65-79.

Hacker, K. L., Freedman, E. G., Gorman, M. E., and Isaacson, R. 1990. The emergence of task representations in small-group simulations of scientific reasoning. *Journal of Social Behavior and Personality* 5:175-186.

Hoenkamp, E. 1989. 'Confirmation bias' in rule discovery and the principle of maximum entropy. In Proceedings of the Eleventh Annual Conference of the Cognitive Science Society, 551-558. Hillsdale, N.J.: Lawrence Erlbaum.

Klahr, D., and Dunbar, K. 1988. Dual space search during scientific reasoning. *Cognitive Science* 12:1-48.

Klahr, D., Dunbar, K., and Fay, A. L. 1990. Designing good experiments to test bad hypotheses. In J. Shrager and P. Langley Eds. *Computational models of scientific discovery and theory formation* (pp. 356-402). Palo Alto, CA: Morgan Kaufmann.

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

Klayman, J., and Ha, Y-W. 1989. Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15:317-330.

Laughlin, P. R., and Futoran, G. C. 1985. Collective induction: Social combination and sequential transition. *Journal of Personality and Social Psychology* 48:608-613.

McDonald, J. 1990. Some situational determinants of hypothesis-testing strategies. *Journal of Experimental Social Psychology* 26:255-274.

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

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

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