Self vs. Other-Generated Hypotheses in Scientific Discovery.[1]

Christian D. Schunn
Department of Psychology
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
schunn@cmu.edu

David Klahr
Department of Psychology
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
klahr@cmu.edu

Abstract

Other-generated hypotheses are often considered easier to test than self-generated hypotheses. To determine the precise effects of other-generated hypotheses, we propose three kinds of effects and describe a study designed to test for these effects of hypothesis source. The three kinds of effects considered are: (i) hypothesis plausibility changes, (ii) skepticism changes, and (iii) process changes. Forty-two undergraduate subjects were given a microworld discovery task called Milktruck. Subjects either had to generate their own initial hypothesis or were given the most frequently generated hypothesis. It was found that the other-generated hypothesis lead to more thorough investigation of hypotheses resulting in a decrease in false terminations with incorrect solutions. The results suggested these effects were caused by an increase in skepticism rather than changes in hypothesis plausibility or process changes.

Introduction

It has been often noted that scientists are more likely to believe hypotheses of their own making than hypotheses proposed by others. Indeed, although we are trained to generate rival hypotheses in all instances, we are usually better able to do so when the hypothesis in question is not one that we have formulated or induced on our own. This general phenomenon raises some important questions about the psychology of scientific discovery. Are the differences in the evaluation of self-generated versus other-generated hypotheses simply a matter of differences in plausibility or skepticism, or are qualitatively different processes invoked in the two situations?

To formulate and address this question precisely, we will examine this question within a framework that views scientific discovery as a coordinated search in two spaces: a space of hypotheses and a space of experiments (Klahr & Dunbar, 1988). We compare the search of two groups of subjects who differ in one critical feature of the discovery process. For one group, the initial hypothesis is the one of their own making. For the other, the initial hypothesis has been suggested by the experimenter (and attributed to another subject). In order to minimize differences between the self-generated and the other-generated conditions, the other-generated hypothesis is the one most likely to have been generated if it hadn't been suggested by the experimenter.

If indeed, the source of an hypothesis can affect the discovery process, then it becomes important to determine precisely the ways in which the discovery process is changed. In the paper we focus on three possible effects:

(i) Plausibility changes. These changes affect features of the way that particular hypotheses (either the specific hypothesis in question, or possibly other hypotheses) are represented. For example, a self-generated hypothesis might be given a higher *a priori* plausibility rating, whereas one provided by someone else might be given a low rating.

(ii) Skepticism changes. These are changes, not in plausibility, but in the parameter settings of the discovery process itself. For example, the hypothesis evaluation procedures presumably have some criterion for deciding when a given hypothesis has sufficient evidence to stop testing it. That is, having an other-generated hypothesis might lead to a higher criterion value, which will lead to more rigorous testing.

(iii) Process changes. Changes in hypothesis source might produce direct changes in the discovery process itself. Such effects might be at the level of strategy changes. For example, a self-generated hypothesis might lead to immediate experimentation, whereas an other-generated hypothesis might cause the researcher to mentally search through the set of possible alternative hypotheses (called hypothesis space search) before beginning experimentation.

In sum, we distinguish between three kinds of effects: data parameter changes (e.g., changes in

[1]This research was funded by scholarships from FCAR and NSERC to the first author, and by grants from the National Institute of Child Health and Human Development (R01-HD25211) and the A.W. Mellon Foundation to the second author.

hypothesis plausibility), process parameter changes (e.g., changes in skepticism), and changes in strategy use (e.g., increase in hypothesis space search).

There is one study which provides indirect evidence for the effects of hypothesis source. Klahr, Dunbar, & Fay (1990) gave subjects hypotheses to test in a computer microworld that were either plausible or implausible, and had subjects make discoveries that were *a priori* either plausible or implausible. Although Klahr et al. did not have a condition in that study in which subject had to generate their own initial hypothesis, they compared their subjects' behavior with the behavior of subjects in an earlier study in which the students had to generate their own initial hypotheses (Klahr & Dunbar, 1988). Klahr et al. noted two main effects of the difference in hypothesis source. First, subjects were more likely to test multiple hypotheses when presented with a hypothesis than when they had to generate the initial hypothesis. Second, subjects were less likely to retain the given hypothesis in the face of negative evidence than if they had generated the hypothesis themselves.

To explain these effects, Klahr et al. claimed that self-generated hypotheses are given higher strength values. In other words, they claimed that it was only the data parameter they called strength (or plausibility) that was effected by the manipulation.

Since Klahr et al. study did not have a direct measure of strength or plausibility, it is unclear whether the effects of hypothesis source did change a data parameter such as strength. Furthermore, Klahr et al. did not have a direct comparison of self-generated vs. other-generated hypothesis conditions, leaving open the possibility that the differences were due to other differences between the Klahr et al. and the Klahr & Dunbar studies.

In the present study, we address some of these potential problems and further investigate the effects of hypothesis source. First, there is a direct comparison, rather than an indirect comparison, of self-generated vs. externally-given hypotheses. Second, subjects are asked to make plausibility judgments for the given and generated hypotheses. The questions to be addressed by the study are: what are the details of the effects of hypothesis source, and are the effects brought about by data parameter changes (such as plausibility), process parameter changes (such as skepticism), or changes in strategies and heuristics?

If the plausibility ratings are affected by the change in hypothesis source, then a data parameter change is implicated. If, however, the plausibility ratings of all hypotheses change, rather than just of the Other-Generated hypothesis, then a process parameter such as skepticism is implicated. Alternatively, if plausibility ratings do not change at all, then strategy and heuristic changes become more likely, and measures of strategy use (e.g., number of hypotheses generated) become important.[2]

Table 1. The function of the arguments of the δ command.

For the last N steps in the program, δ reorders the execution sequence of the program by...	
◿ (increasing)	...increasing house number order.
◺ (decreasing)	...decreasing house number order.

Method

Subjects. Forty-two Carnegie Mellon University undergraduates took part in the experiment for course credit. All of the subjects had used a computer before, and over 85% of the subjects had some programming experience.

The computer interface. Subjects worked in a complex microworld called Milktruck (Schunn & Klahr, 1992). In the microworld, a "milk truck" executed a sequence of actions associated with a dairy delivery route. At any of 6 different locations along its route, it could beep its horn, deliver milk or eggs, or receive money or empties. The route of the milk truck was programmed used the keypad. A program consisted of a sequence of action-location pairs. As subjects entered their programs, the steps were displayed on the screen in the program listing. After the route had been entered, the subject pressed 'RUN' and the milk truck executed its route on the screen. The milk truck went to each location on the programmed route in the order that it was programmed, and the subjects were shown by way of animated icons what transpired at the location. Also, as the route was completed, a trace listing displayed in program format what transpired during the run (see figure 1).

When the mystery command, δ (delta), was not used, the trace listing was identical to the program listing. However, the δ command could change the order of delivery, and the resultant trace would then be discrepant from the program listing. The effect of the δ command was to reorder the execution of part of the program according to the values of its two arguments, a number (1-6), and a triangle (white or black). Table 1 describes the effects of the δ command. Programs 2 and 3 in figure 5 demonstrate the effects of the δ with white triangle and with black triangle respectively.

Procedure. Subjects took part in a single, one hour session. Following an introduction to the basics of the

[2]However, the possibility remains that data and system parameters other than plausibility and skepticism were affected.

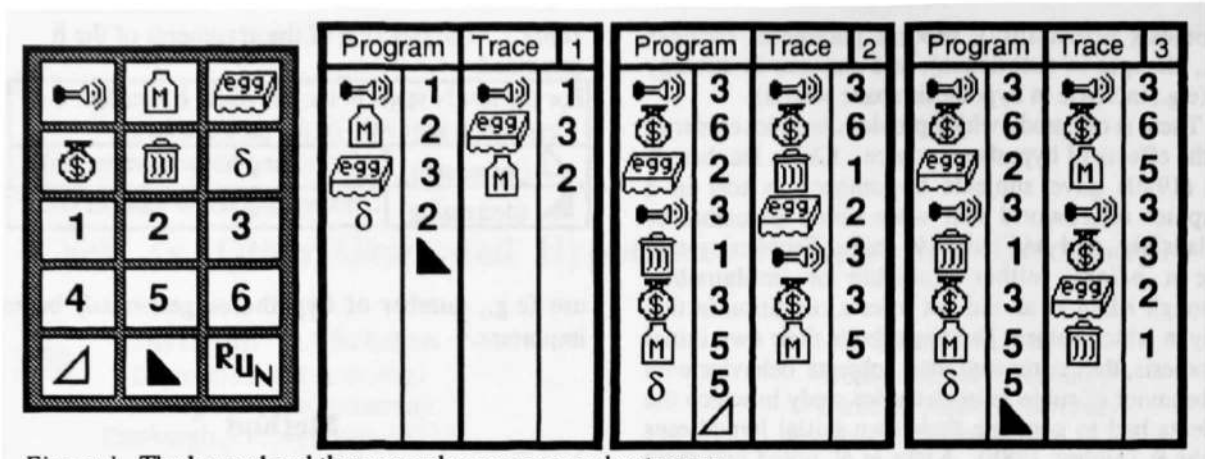

Figure 1. The keypad and three sample programs and outcomes.

Milktruck domain, the syntax of the δ command was described, and the goal of discovering the effect of the δ command was presented to the subjects.

First, subjects were told, "As an example of how to use the delta command, let's try entering a program tried by the previous person in the experiment." Second, all subjects were given the same example program, which is shown as the first program in figure 1. This program was chosen so that a large number of simple hypotheses would be compatible with the outcome.

At this point the experimental manipulation was introduced. When the program had finished running, the subjects were given different conclusions about the outcome of the program. Subjects in the Other-Generated condition were told that the previous subject had concluded that the δ puts the delivery at the house number equal to the number after δ last. This hypothesis was the most frequently generated initial hypothesis in the Self-Generated condition.[3] Subjects in the Self-Generated condition were not given an initial hypothesis.

All subjects were then asked to generate as many hypotheses as they could about how the δ key worked, and to give a plausibility rating for each of the generated hypotheses. Following the hypothesis generation task, the subjects went on to the discovery phase. Here, subjects designed, conducted, and analyzed experiments with the goal of discovering the role of the δ command and its arguments. The subjects worked at the discovery task until they felt that they had solved it, or 40 minutes had elapsed. Subjects were free to terminate with incorrect hypotheses.

Measures and predictions. There are four kinds of data generated by this experiment. First, the final solution rates provide an evaluation of the importance of the source of the hypothesis. That is, does changing the hypothesis source appreciably affect success in this task? The Klahr et al. results suggest that there should be more solvers in the Other-Generated condition.

Second, there is the plausibility of the initial hypotheses generated by the subjects prior to the experimentation phase. The plausibility ratings provide an important test of the mechanism by which hypothesis source impacts on discovery behavior. There are several potentially informative outcomes: the plausibility of the initial hypotheses are unaffected by the manipulation; only the plausibility of the Other-Generated hypothesis changes in plausibility, implicating a data parameter effect; or all of the initially-generated hypotheses change in plausibility, implicating a process parameter effect.

Third, there is the number of the initial hypotheses generated by the subjects prior to the experimentation phase. The Klahr et al. results suggest that there may be more hypotheses generated in the Other-Generated condition. If there is no effect of the manipulation on the number of hypotheses generated, then it is unlikely that the manipulation provided any new content about possible hypotheses to the subjects. Such a result would provide a good manipulation check that only hypothesis source, and not hypothesis content, was varied.

Fourth, there is the experimentation phase behavior. Analysis of the number of programs generated, the rate of program generation and interpretation, the informational content of the programs generated, and the strategies used in program generation may reveal whether their were effects of the manipulation at the level of strategies and heuristics. The Klahr et al. results suggest that there should not be any qualitative differences in strategy use between the two conditions.

[3]Over 85% of the subjects in the Self-Generated condition generated this hypothesis prior to the experimentation phase. The remaining subjects either generated close variants or gave ambiguous responses.

Results

Final outcomes. Three solution groups will be used for the analyses: subjects who came to the correct solution (Solve); subjects who self-terminated with an incorrect solution (False Solution); and subjects who reached the time limit without solving (Not Solve). Table 2 presents the number of subjects in each solution group in each of the discovery conditions.

To assess the effects of condition on solution group, ANOVA's were computed on the proportion of Solvers and False Solutions in the two conditions. The greater proportion of solvers in the Other-Generated condition was not significant ($F(1,40)<1$); whereas the smaller proportion of False Solutions in the Other-Generated condition was significant ($F(1,40)=4.6$, $p<.04$).

Initial Hypotheses. An ANOVA conducted on the rated plausibility of the hypothesis given in the Other-Generated condition revealed that the manipulation produced only a marginally significant difference in the plausibility of that hypothesis ($F(1,27)=2.1$, $p<.2$). Table 6 presents the mean rating in each condition.

While the plausibility of the given hypothesis was not strongly effected by the manipulation, an ANOVA conducted on the mean plausibility of the first four hypotheses generated by each subject revealed that there was an overall decrease in rated plausibility of hypotheses in the Other-Generated condition ($F(1,41)=6.3$, $p<.02$). Table 3 presents the means for each condition. Taken with the weak effect on the plausibility of the given hypothesis, these results suggest that there was a change in skepticism (a process parameter), rather than a change in plausibility (a data parameter).

An ANOVA conducted on the total number of hypotheses generated (including the given hypothesis) revealed that being given an initial hypothesis did not significantly increase the number of hypotheses generated ($F(1, 41)=1.1$, $p>.3$). Table 3 presents the means for each condition. These results suggest that manipulation did not change the amount of information that the subjects had about possible hypotheses.

Experimentation phase behavior. To test whether the manipulation changed the amount of time that the subjects spent on the task, ANOVA's were conducted on the number of programs run, the total time spent conducting programs, and the time spent per program (see table 4 for the means). The subjects in the Other-Generated condition ran significantly more programs ($F(1,39)=11.6$, $p<.002$), over significantly more time ($F(1,39)=4.76$, $p<.04$). However, there was no effect on the amount of time spent on each program ($F(1,39)=1.5$, $p>.2$). These results suggest that the manipulation caused subjects to test their hypotheses more rigorously. Did subjects in the Other-Generated condition produce more informative programs? One simple measure of the information content of an experimental outcome is the number of changes between the program and trace listing. For example, the second program in figure 1 consists of 3 changes[5]. An ANOVA on the total program-trace changes for each subject was computed. Table 5 presents the means for each condition. The greater number of program-trace changes in the Other-Generated condition was significant ($F(1, 39)=12.5$, $p<.001$).

This greater total information content may merely reflect the greater number of programs run. To test this

Table 2. The number of subjects in each solution group in each of the discovery conditions.[4] *=p<.05

	Self-Generated	Other-Generated	
Solve	6 (25%)	6 (33%)	
False Solution	13 (54%)	4 (22%)	*
Not Solve	5 (21%)	8 (44%)	
Total	24 (100%)	18 (100%)	

Table 3. Mean plausibility of the given hypothesis, the mean plausibility of the first four initial hypotheses, and the mean number of initial hypotheses in each condition (with standard deviations). **=p<.02, *=p<.2.

	Self-Generated	Other-Generated	
Plausibility of Other-Generated hypothesis	61.2 (22.6)	48.2 (25.4)	*
Mean plaus. of first four initial hypotheses	56.9 (21.2)	42.2 (15.9)	**
Number of initial hypotheses	4.3 (1.6)	4.8 (2.0)	

Table 4. The mean number of programs, time on task, and time per program in each of the conditions (with standard deviations). **=p<.01, *=p<.05.

	Self-Generated	Other-Generated	
# of programs	16 (7.5)	24.6 (8.5)	**
Min. on task	24.7 (12.6)	32.5 (9.3)	*
Min. per prog.	1.6 (.57)	1.4 (.39)	

[4]Random assignment produced strong imbalances in several individual difference measures which proved to be very predictive of final solution (overall r=.43). Extra subjects were run to reduce these imbalances across conditions, thereby resulting in unequal numbers of subjects across conditions.

[5]From the verbal protocols, it was noticed that subjects tended to focus on the changes made in the program. Therefore this measure of information content was used rather than some arguably more crucial measure involving program length and the number after δ.

hypothesis, an ANOVA was computed on the number of program-trace changes per program. Table 5 presents the mean number of changes per program for each condition. The small difference shown in the table is not significant ($F(1, 39)<1$) indicating that the subjects in the Other-Generated condition were not more effective at producing informative programs, according to this measure of information content.

However, it may be that there are critically informative programs (e.g., programs with many changes). In such a case, a better measure of the information content generated by a subject would be of the maximally informative program. To test whether the manipulation influenced the ability to generate maximally informative programs, an ANOVA was calculated on the maximum number of program--trace changes generated by each subject. Table 5 presents the mean maximum for each subject in each condition. The mean of the Other-Generated condition was marginally higher than that of the Self-Generated condition ($F(1, 39)=2.9$, $p<.1$).

Yet, since the subjects in the Other-Generated condition ran more experiments, it may be that simply by chance that their maximum number of changes is higher. Furthermore, more changes tend to be produced later in the task when subjects are using longer programs. Therefore, the maximum number of program changes for each subject in the Other-Generated condition was recalculated using only the first 65% of their programs, since the subjects in the other condition ran only 65% as many programs. Table 5 presents the means of this adjusted maximum. As expected, there were no differences across conditions ($F(1, 39)<1$), suggesting that the subjects in the two conditions did not differ in their ability to produce informative programs.

Another measure of program information content can be derived from an analysis of which regions of the Experiment space were investigated by the subjects. Analyses by Klahr & Dunbar (1988) suggest that certain combinations of the number argument to δ (N) and the program length (λ) are important variables to explore in order to discover this particular role of N. The region of importance is the one with large values of both N and λ, since long programs with a large number argument to δ are more likely to generate many program changes.[6]

To investigate the effects of the manipulation on this variable ANOVA's was computed on both the number and proportion of programs with N>3 and $\lambda>3$.[7] The mean number and proportion of programs in this region are presented in table 6. Neither of the effects were statistically significant ($F(1, 39)=1.1$, $p>.3$, and $F(1, 39)<1$ respectively).

Table 5. Mean number of total program-trace changes, mean number of changes per program, the mean of the maximum program-trace changes for each subject, and the mean of the maximum number of program changes for each subject adjusted for total number of programs (with standard deviations). **=p<.01, *=p<.1.

	Self-Generated	Other-Generated	
Total program-trace changes	21.6 (13.3)	36.2 (13.0)	**
Mean changes per program	1.36 (0.64)	1.52 (0.44)	
Mean maximum change	4.1 (1.1)	4.7 (0.9)	*
Mean maximum (equal programs)	4.1 (1.1)	4.1 (1.2)	

Table 6. Mean number of programs with N>3 and $\lambda>3$, and the mean proportion of programs with N>3 and $\lambda>3$ (with standard deviations).

	Self-Generated	Other-Generated
# of programs with N>3, $\lambda>3$	4.0 (3.4)	5.1 (3.4)
Proportion of prog with N>3, $\lambda>3$	0.26 (0.17)	0.23 (0.15)

Discussion

Changes in the source of a plausible initial hypothesis produced several effects: a decrease in the number of false solutions, an increase in the number of programs run, an increase in the time on task, an increase in the total information content of the programs, and an overall decrease in rated plausibility of the initially generated hypotheses. These effects suggest that the manipulation produced increased skepticism in the subjects, causing the subjects to go beyond their usual termination criteria and test hypotheses with greater rigor.

However, there is no evidence that the manipulation caused changes in the discovery strategies and heuristics—measures of Hypothesis space strategies and heuristics (e.g., number of hypotheses generated) and Experiment space strategies and heuristics (e.g., time spent per program, the amount of information produced per program, the regions of the Experiment space that were visited) did not indicate any qualitative changes.

That the plausibility of the given hypothesis did not drop significantly, and that there was an overall drop in plausibility of all hypotheses suggests that

[6] The correlation between program length and program changes is .39 across all subjects ($F(1,799)=142$, $p<.0001$).

[7] This cutoff point was chosen somewhat arbitrarily. Selected segments of length less than or equal to three often produce results which are highly ambiguous.

effects were not at the level of a data parameter such as plausibility; rather the results are more consistent with a process parameter change such as an increase in skepticism.

At an abstract level, the effects are both interesting and surprising. In the control condition, subjects were placed in an environment in which they generated and endorsed the initial hypothesis A (deliver to house N last). The structure of the actual role of the mystery function was such that this initial hypothesis A was disconfirmed very early in the discovery process. Many of the subjects then induced hypothesis B (reverse last N steps), and terminated falsely with this final hypothesis. The correct hypothesis was in fact hypothesis C (sort the last N steps). It was by externally giving subjects the initial hypothesis A, the same hypothesis that they would have generated anyway, that caused subjects to later go beyond hypothesis B.

The manipulation was fairly subtle. Subjects were not given a set of hypotheses to test, as is often done (e.g., Sodian, Zaitchik, & Carey, 1991; Klahr, Fay, & Dunbar, 1993). Rather, subjects were merely told that another subject presented with the same information had generated this hypothesis, thereby manipulating the source of the hypothesis and not the discovery task given to the subjects.

Also, the effects of the manipulation could not be at the level of providing information to the subjects, either explicitly (by providing a new hypothesis) or implicitly (by endorsing or make dubious a hypothesis), since the Other-Generated hypothesis was one that the subjects could and did generate easily, and this hypothesis was discarded by all subjects early in the task.

It is important not to confuse the effects of hypothesis source with the effects of having alternative hypotheses. Some might argue that scientists are always in the Other-Generated condition since colleagues are constantly providing alternative hypotheses. By this argument, the self-generated condition may be considered an artificial one. However, since primary hypotheses can be self-generated and other-generated, and alternative hypotheses can be self-generated and other-generated, the issue of alternative hypotheses is orthogonal to the issue of hypothesis source. Further research is required to investigate possible interactions in the effects of these two factors.

Contrary to the Klahr et al. findings, this manipulation did not produce an increase in the number of hypotheses generated. This difference in findings may be accounted for by the differences in the procedures by which subjects were asked to produce initial hypotheses. In this task, subjects were given several probes designed to evoke as many different hypotheses as possible. As a result, this procedure may have produced a ceiling effect.

Interestingly, findings apparently contradicting our results were reported by Koehler (1992), who found that self-generated hypotheses were given *lower* confidence ratings than other-generated hypotheses. In a series of experiments, Koehler presented evidence which suggested that the self-generated hypotheses were given lower confidence ratings because the process of hypothesis generation forces the individual to consider more alternative hypotheses. However, many potentially important factors differ between the task given to Koehler's subjects and the task used in the current study. For example, the task used by Koehler was a pure hypothesis generation task rather than a discovery task. Therefore, it is unclear whether the same processes are being used in the two tasks. Further experimentation is required to identify which of the factors that differ between the two domains of study is responsible for these apparently contradictory results.

Of course, hypothesis source is not the only important factor involved in determining the acceptance of a hypothesis. However, this research has shown that hypothesis source can be an important factor. Additional experimentation using similar kinds of explicit measures of possible mediating factors such as hypothesis plausibility and skepticism are required to further investigate the path by which hypothesis source leads to changes in the discovery process.

References

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

Klahr, D., Dunbar, K., & Fay, A. 1990. Designing Good Experiments to Test "Bad" Hypotheses. In J. Shrager & P. Langley (Eds.), *Computational Models of Discovery and Theory Formation*. Hillsdale: LEA.

Klahr, D., Fay, A. L., & Dunbar, K. 1993. Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, **25**, 111-146.

Koehler, D. J. 1992. Generating Your Own Hypotheses Makes You Less Confident It's True. Poster presented at the 4th Annual Convention of the American Psychological Society.

Schunn, C. D., & Klahr, D. 1992. Complexity Management in a Discovery Task. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, 177-182. Cambridge, MA: MIT Press.

Sodian, B., Zaitchik, D., & Carey, S. 1991. Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, **62**, 753-766.