Exploration in the Experiment Space: The Relationship between Systematicity and Performance

Susan B. Trickett (stricket@osf1.gmu.edu)
George Mason University, Fairfax, VA 22030

J. Gregory Trafton (trafton@itd.nrl.navy.mil)
NRL, Code 5513, Washington, DC 20375

Paula D. Raymond (raymond@xp.psych.nyu.edu)
Center for Applied Research

Abstract

Much of the research on scientific reasoning has investigated the use of explicit, hypothesis-testing strategies. However, there is evidence that scientific reasoning problems can be solved by exploration of the experiment space. This study investigates the strategies by which people explore the experiment space and examines the relationship between the systematicity of this search and successful performance.

Introduction

The processes underlying successful performance on scientific reasoning tasks represent an important skill, or set of skills, that are crucial for students of science at all levels in all disciplines to acquire. However, the precise nature of these skills, as well as the extent to which they might transfer across scientific domains, is still an open question.

One of the primary theoretical frameworks in which scientific reasoning has been studied is Klahr and Dunbar's (1988) characterization of scientific reasoning as a search in two problem spaces, a hypothesis space and an experiment space. Although recently the exact number and nature of the search spaces has become a matter for debate (Baker & Dunbar, 1996; Schunn & Klahr, 1995) the distinction between hypothesis-formation and experimentation has remained.

Klahr and Dunbar (1988) analyzed participants' performance in figuring out how the repeat key worked on a programmable robot (BigTrak). They identified two types of participants, characterized by differences in the way they searched the different spaces. Participants working in the hypothesis space (called theorists) were able to form new hypotheses by searching memory. They stated and tested hypotheses more frequently than participants who worked in the experiment space. On the other hand, participants working in the experiment space (called experimenters) stated fewer explicit hypotheses, conducted more experiments, and took longer to find the solution. Furthermore, they formed the correct hypothesis only as a result of running experiments. Interestingly, however, the experimenters were ultimately as successful in solving the problem as the theorists.

Most research on scientific reasoning has focused on explicit hypothesis-testing strategies and is thus associated with performance in the hypothesis space. In order to understand what differentiates successful from unsuccessful performance, many researchers have examined the explicit hypothesis-testing strategies by which people attempt to solve scientific reasoning tasks. Several such strategies have been identified. The varying-one-thing-at-a time strategy, or VOTAT (Tschirgi, 1980) involves holding all variables constant except one. The Change-All strategy (Tschirgi) involves changing the value of every variable. People who use the engineering strategy (Schauble, Klopfer, & Raghavan, 1991) try to bring the system to a particular desirable state. The holding-one-thing-at-a-time, or HOTAT (Tschirgi) strategy involves holding one variable constant and changing all others. The confirmation strategy (Wason, 1960) involves proposing a hypothesis and seeking to confirm, rather than disconfirm, it.

Of the strategies outlined above, VOTAT is the only consistently effective hypothesis-testing strategy. It has been closely associated with successful performance (Vollmeyer et al. 1996; Shute & Glaser, 1990). For example, Shute and Glaser found that in using an economics microworld to "discover" the laws of supply and demand, more successful participants typically changed only one variable at a time. Vollmeyer et al. had participants discover the relationships among variables in a biology task. As participants shifted away from the change-all strategy to the VOTAT strategy, the number of correct answers increased.

The hypothesis-testing strategy people use to solve this type of task is thus an important part of successful performance. As Vollmeyer et al. (1996) explain, VOTAT is an effective strategy, because it "allows the logical disconfirmation of alternative hypotheses." However, many of the experiments conducted by the experimenters in Klahr and Dunbar's (1988) BigTrak study were not designed to test an explicit hypothesis, and yet these participants were still able to solve the problem correctly. This suggests there may be a distinction between their data collection and their interpretative strategies. Tschirgi (1980) also suggested that adults, as well as children, manifest this same separation. She proposes that people do not necessarily analyze the logical underpinnings of their experimentation.

Not much research has explored the characteristics of performance in the experiment space, yet a number of questions arise. When people explore the experiment space without explicitly using a hypothesis-testing strategy, how are they able to solve scientific reasoning problems successfully? Are there patterns of behavior in gathering data that are associated with being able to interpret and explain those data appropriately? There is some evidence that this might be the

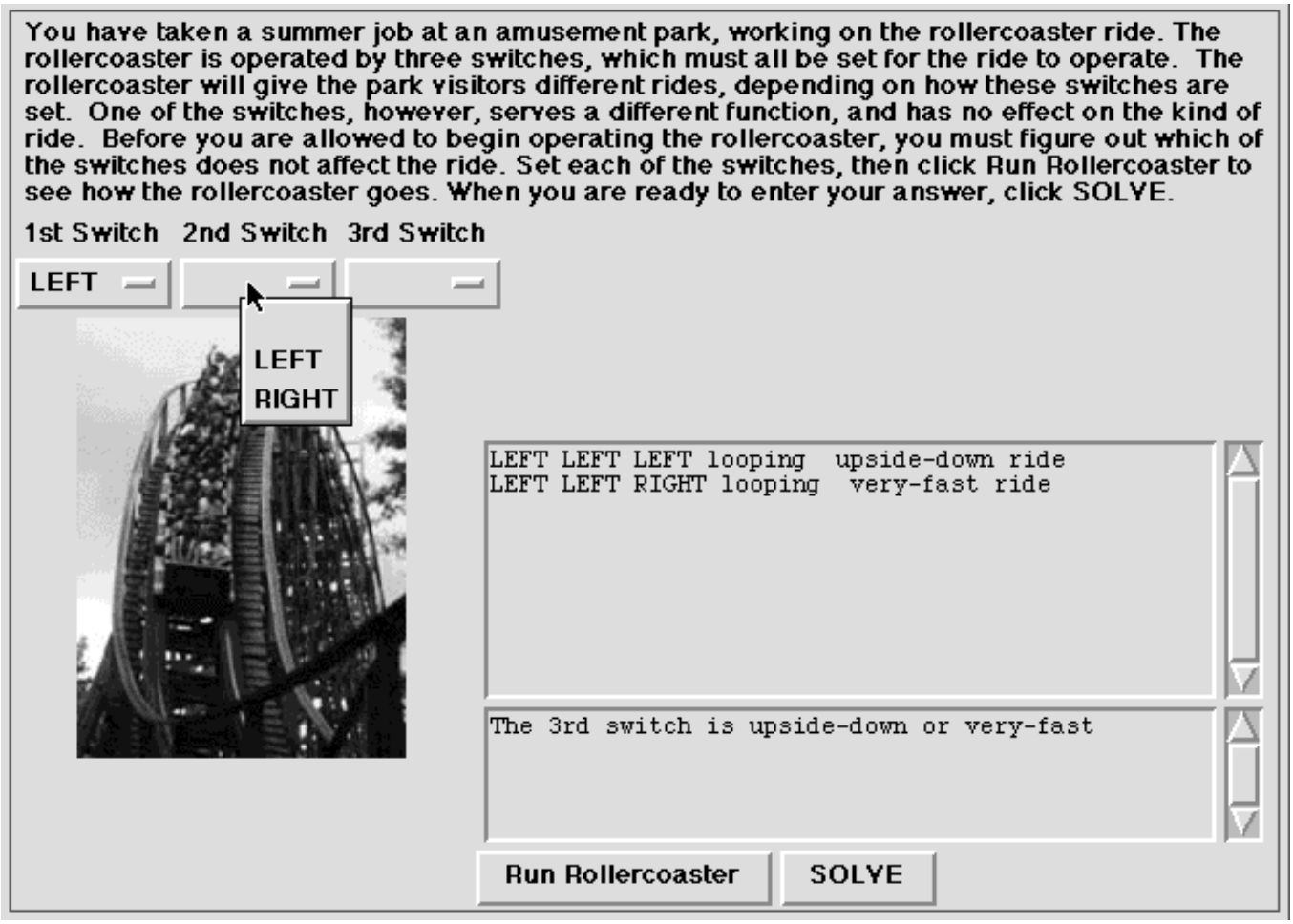

Figure 1: Screenshot of interface

case; for example, Shute, Glaser, and Raghavan (1989) analyzed performance of two successful and two unsuccessful participants in Smithtown, an economics microworld, and found that in addition to effective thinking and planning, efficiency in data management was associated with more successful performance. Their results suggest that the method by which people collect data is indeed important.

The present study explores the relationship between exploration in the experiment space and performance. Specifically, we wanted to determine whether data collection skills are separate from the hypothesis-testing strategies discussed above and to determine whether systematic data collection is related to successful performance.

We chose a task for which the experiment space was constrained and quite small, so that participants would not find the task too complex to solve. Pilot tests indicated that although the task involved only three variables, solving it was not trivial for our participants. The small experiment space has the further advantage of making it easier to determine the systematicity with which participants collect their data.

We also wanted to explore how these data collection strategies are used when participants solve the same task more than one time and whether they transfer from one task to another. Consequently, there were two conditions in the study, an identical task condition and an isomorphic task condition, to study strategy transfer.

Method

Participants

Participants were 30 George Mason University undergraduates (16 males and 14 females), who received course credit for their participation. Participants were randomly assigned to one of two conditions. Protocol data from five participants was lost because of equipment failures; data from these participants was not included in the analysis.

Materials

Five isomorphic tasks were developed, based on an adaptation of a task from Siegler and Atlas (1976). As an example, in one task participants were asked to solve a problem about a rollercoaster. The rollercoaster was operated by three switches and gave a different ride, depending on how the switches were set. Although all three switches had to be set for the rollercoaster to work, one of the switches did not affect the kind of ride. The task was to identify which switch did not affect the rollercoaster ride. Each switch had two pos-

sible settings (left or right); participants had to select a setting for each of the three switches, then run the rollercoaster. The resulting ride was represented on the computer screen.

Participants could run as many tests (trials) as they wished, and the interface included a "notepad" on which they could type comments if they chose. Figure 1 shows a screen snapshot of the rollercoaster task. The interface was the same for each task; only the instructions, variables, and answer were different across tasks. We used the same interface to avoid any interaction between performance and interface.

A different cover story was developed for four additional tasks in different domains. The tasks were isomorphic in that they shared the same deep structure and could be solved by applying identical procedures (Simon & Hayes, 1976). For each task, there were three possible causal variables, each with two settings. One variable had no effect, and the task was to identify that variable. The optimal solution strategy in each task was to test each variable by changing its setting, while holding the other variables constant. If the result was the same in both test, one could deduce that the variable had no effect. If the result was different, one could infer that that variable *did* have an effect and test another variable in the same way.

Design

There were two conditions in this study: a "same task" condition, (hereafter referred to as the identical condition), and an "isomorph" condition. There were five different tasks. In the identical condition, participants were asked to solve the same task five times. Each of these isomorphs was given to three participants in this condition. In the isomorph condition, participants were asked to solve a series of five isomorphic tasks, one each of the five tasks used in the identical condition. The tasks were partially counterbalanced, so that each task appeared three times in the last position of the sequence. The correct solution for each task was randomly generated in both conditions.

Measures

Keystroke data was collected and time-stamped as participants solved the tasks. In addition, verbal protocols of the entire task sequence were collected (Ericsson & Simon, 1984). Verbal protocols were coded as described below (note that only those measures relevant to the present analysis and discussion are included here).

Two measures of solution correctness were used. The first was generated from the keystroke data and measured only whether the solution entered was correct (right answer). Because of the high probability (.33) that a correct answer could be chosen by chance, a second measure (right answer/right reasoning) was developed using the verbal protocols. This measure identified correct solutions for which there was evidence of explicitly verbalized correct reasoning.

In order to investigate the participants' use of the hypothesis space, verbal protocols were also coded for explicit hypothesis-testing strategies used, specifically whether the VOTAT strategy was adopted. A participant was coded as using the VOTAT strategy only if he or she explicitly verbalized this strategy. In addition, protocols were coded for identification of an a priori plan.

An important component of the task was recognizing that the results of two different tests were the same. We used the verbal protocols to code whether participants noticed this, and if so, on what trial they noticed it. Table 1 shows the coding scheme used, with examples from the protocols.

Table 1: Coding scheme

ID	Code	Utterance
205	right answer/ right reason	Bronze, because the first 2 are the same, the bronze is different, and it doesn't change
107	VOTAT	I just have to keep 2 the same and change one and see if it's different
203	plan	So I need to test the strings
204	notice-2-same	It's the same ride

In order to investigate the strategies by which participants explored the experiment space, protocols were coded for systematicity of data collection. There were eight possible combinations of variables that participants could select and test. They could run as few or as many tests as they chose, and tests could be repeated or duplicated. For data collection to be coded as systematic, 75% or more of the tests chosen had to conform to a discernible pattern.

left left left	left left left
left left right	left left right
left right left	left right left
left right right	right left left
right left left	right right right
right left right	
right right left	
right right right	

Figure 2: Systematic strategies

Several different patterns of data collection emerged. Figure 2 shows two examples of systematic data collection strategies. Some participants recognized that there were eight possible combinations and organized their data by conducting all four tests with one variable at one setting before conducting a second block of four tests with that variable at its second setting. Frequently, the second block of tests was ordered in exactly the same way as the first. Some participants changed each variable in turn, while keeping the other two the same. (Note that this was different from the VOTAT strategy in two ways; first, participants did not explicitly state that they were using VOTAT, and second, participants frequently continued to run more tests after changing each of the variables. If they had been using VOTAT, an explicit hypothesis-testing strategy, they would not presumably have continued with these extraneous tests). Both the strategies described above were coded as systematic.

Some participants simply tried various combinations without close attention to prior tests. Such a strategy was characterized by duplications of tests which frequently went unnoticed. Others attempted to find new combinations and avoid duplications on an ad hoc basis (i.e., trial by trial).

Others realized there were eight combinations and tried to find all of them, again, on an ad hoc basis. These two last approaches were characterized by participants searching the records of prior tests to see if they had already tried a given combination. In the last approach, participants searched the data to see if they had tested all the combinations. The strategies described above were coded as unsystematic.

To date, only one coder has coded the protocols; consequently, no inter-rater reliability scores will be reported.

Procedure

Participants were trained on the talk-aloud process and on the features of the interface used for the experimental task (i.e., how to set variables, run tests, interpret the results, and use the notepad). When participants understood the interface, they began the experimental task. They proceeded through the series of five tasks without receiving feedback as to whether their solutions were correct.

Results and Discussion

We first analyzed the time it took participants in each condition to complete the series of five tasks. The mean completion time for the identical condition was 23 minutes 44 seconds, and for the isomorph condition it was 30 minutes 20 seconds. It is most likely that participants in the identical condition took less time to solve the task because they did not need to re-read the task instructions every time. However, in general these times are not very informative, because we were collecting verbal protocols.

Correct Solution

We analyzed performance, measured by the number of correct solutions according to the keystroke data. The identical condition showed little if any improvement; means for tasks 1 through 5 were, respectively, .72, .45, .45, .63, .63 (N = 11). In the isomorph condition, performance improved, shown by an increase in the mean correct. Respective means for tasks 1 through 5 in the isomorph condition were .64, .78, .71, .92, and .92 (N = 14).

Because of the high probability of a correct answer's being chosen by chance, we analyzed performance using the right answer/right reason measure described above. Note that all further analyses of correctness use this measure.

Again, in the identical condition, performance did not improve; means for tasks 1 through 5 were .45, .36, .27, .36, and .45, respectively (N = 11). In the isomorph condition performance improved as participants progressed through the tasks; means correct were .29, .50, .42, .79, and .71, respectively (N = 14). A test for increasing linearity showed that for participants in the isomorph condition, performance improved in a linear fashion as they progressed through the series of five tasks, F (1, 13) = 13.42, MS_E = .16, p < .05.

Why didn't performance in the identical condition improve? The protocols suggest two possible reasons. First, because there was no feedback, some participants thought that when the same task appeared it meant they had got the answer wrong. Some participants did not recognize that the tasks were independent of one another and consequently tried to carry information from one task over into another. For example, if they chose the first variable on the first task, they thought that the first variable was not the correct choice for the next task, regardless of what the data indicated.

Second, some participants in this condition began with a poor representation of the task. They never revised this representation, and so they kept getting the answer wrong. On the other hand, if a participant began with an appropriate representation of the task, he or she could begin by solving the task correctly and would then continue with correct performance. In either case, there would not be an *increase* in the number of correct solutions.

If this were the case, we would expect the majority of participants in the identical condition to have got almost all the tasks either incorrect or correct. In fact, the participants who either got four or five of the tasks wrong or got all five of the tasks right account for over 80% of the participants in the identical condition. It would appear then that either these participants started out well and continued well or they started badly, got stuck in a rut, and perseverated on their strategy because the task did not change. Participants in the isomorph condition did not have these problems. The lack of feedback did not affect them in the same way because after they had entered their answer, they got a different task and did not associate the new task with their performance on the previous one. Furthermore, with each new task, they had to construct a new representation. If they did construct a poor representation, it affected performance on that task only.

Systematicity

Performance in the isomorph condition improved as participants progressed through the sequence of tasks. What caused this improvement? Possibly, participants recognized the isomorphism between tasks and used that information to solve the problems more accurately. In fact, only three participants made reference to similarities to previous tasks. Perhaps they came to a better understanding of the task and developed a better strategy to solve the problems. To explore the role of planning and strategy use in performance on these tasks, we examined whether participants stated an a priori plan and whether they used the VOTAT strategy. We also examined the systematicity with which they collected data in these tasks.

Explicit Hypothesis-Testing Strategy Use and Planning Recall that explicit strategies and plans were coded only if participants verbalized their use. Only three participants stated an a priori plan. Overall use of the VOTAT strategy was also very small; only three participants explicitly used this strategy. In the identical condition, only one participant used this strategy, and this use was on the last task in the series. In the isomorphic condition two participants used the VOTAT strategy, one on the fourth task and both on the last task. Similarly, there were very few instances of other explicit strategies (engineering, HOTAT, and confirmation).

Thus, although it appears that there was a very small increase in explicit use of the VOTAT strategy, this shift is not sufficient to account for the improvement in performance in the isomorph condition. Apparently, very few of the participants were searching the hypothesis space in order to

solve these problems. If participants were not using an explicit hypothesis-testing strategy, how did they set about solving the tasks? In order to answer this question, we investigated the strategies by which they explored the experiment space.

Data Collection Systematicity Recall that 75% of the trials on a task had to be systematic for performance on that task to be coded as systematic. In the identical condition, there was little if any difference across tasks on this measure. Means for tasks 1 through 5 were .27, .56, .38, .40, and .40, respectively. In the isomorph condition, however, participants became more systematic. Means for this group for tasks 1 through 5 were .29, .50, .33, .86, and .86, respectively. Next we explored the relationship between systematicity and performance in both conditions.

Systematicity and Correctness

Figure 3 illustrates the trends for both correctness and systematicity for each condition. It shows that as participants in the isomorph condition became more systematic their performance became more correct. It also shows that participants in the identical condition did not become more systematic and their performance did not improve. Overall, systematicity was strongly correlated with correct solution, $r = .68$, $p < .01$.

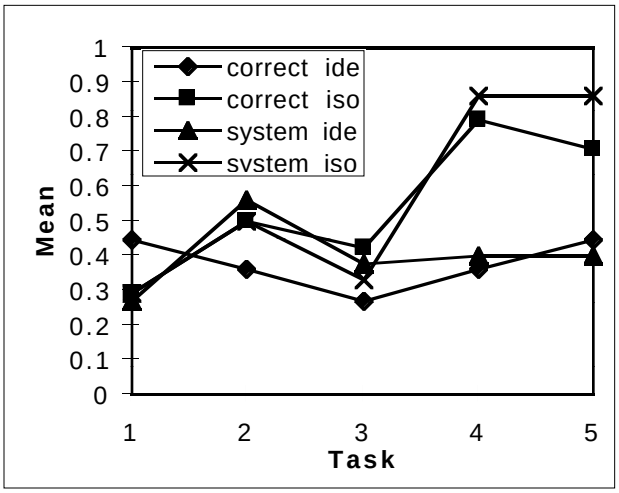

Figure 3: Systematicity and correctness

How did participants in the isomorph condition progress from being unsystematic and wrong in task 1 (mean correct was .29, mean systematicity was .29) to being systematic and right in task 5 (mean correct was .71, mean systematicity was .86)? Why did participants in the identical condition *not* show this progression?

It appears that participants in the isomorph condition gradually refined their systematicity and also became more correct. Figure 4 illustrates this progression. On each task, each participant's performance was coded as non-systematic and not right, non-systematic and right, systematic and not-right, or systematic and right. Three participants in each condition were systematic and right on task 1 and continued this performance on all five tasks. In both conditions, the majority of participants began by being non-systematic and not-right. In the identical condition, performance remained relatively stable. In the isomorph condition, however, a clear shift occurred away from non-systematic, not-right performance to systematic and right. In tasks 4 and 5, no participants in the isomorph condition were non-systematic and not-right, and the majority were both systematic and right.

What could account for the very different pattern in the identical condition? Over 50% of the participants in both conditions began by being non-systematic and not right. Figure 3 shows an initial increase in mean systematicity (from .27 in task 1 to .55 in task 2); however, this increase represents a shift to a systematic strategy by only two participants. Of these two, one solved the problem correctly, and then became unsystematic and wrong for the remaining three tasks. The protocol reveals that after the second task, this participant reverted to her initial poor representation of the task. The other participant shifted to a systematic data collection strategy in task 2 but did not solve the task correctly. The protocol shows that this participant's answers were significantly influenced by his belief that his prior solutions must have been wrong.

What advantage did being systematic in exploring the experiment space gain for participants in the isomorph condition? One of the keys to successfully solving the tasks was noticing that the results of two tests were the same, even though the setting of one variable was different in these two tests. Some participants noticed this immediately, some participants noticed it only after conducting more tests, and some participants apparently did not notice it at all.

Were participants more likely to notice this important piece of information if their data were organized in a systematic way? We analyzed the correlation between systematicity and participants' noticing (immediately or after conducting more tests) that the results of two tests were the same. This correlation was significant, $r = .59$, $p = .005$. This correlation suggests that as participants were more systematic, they were more likely to detect important and relevant patterns in the data, allowing them to successfully solve the problems.

It appears then that there was a strong relationship between being systematic in collecting data and successful performance in the isomorph condition. As participants became more systematic, their performance became more correct. Participants in the isomorph condition began the series of tasks without a systematic data collection strategy and the majority did not solve the task correctly at first. However, by the last task, the majority of these participants were using a systematic strategy *and* were getting the answer right.

Finally, the overall correlation between systematicity and correct performance was positive and significant.

General Discussion

The performance of participants in the isomorph condition improved as they progressed through the series of five tasks. Yet the vast majority of these participants were not using explicit hypothesis-testing strategies such as VOTAT that have been associated with successful performance. Furthermore, hardly any of these participants stated a plan for solv-

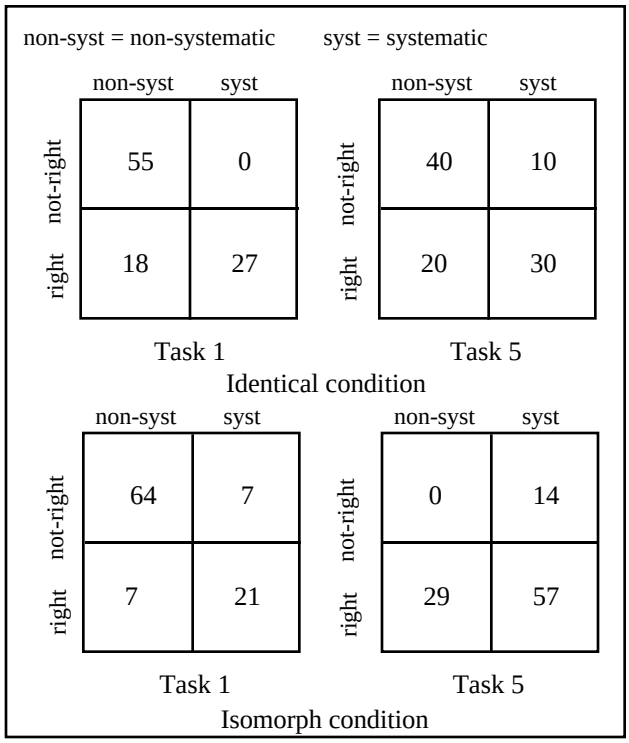

Figure 4: Relation between systematicity and correctness

ing the task. They appear to have been operating almost exclusively in the experiment space.

What can account for this improvement in performance? Participants' improvement occurred primarily in conjunction with their becoming more systematic. It appears then that searching the experiment space can lead to the correct solution, and that what differentiates successful from unsuccessful performance in this space is the systematicity with which the search is conducted. Search in the hypothesis space is more efficient than search in the experiment space; however, it would also appear to be less common and perhaps more difficult. The results of this study suggest that by conducting a search in the experiment space in a systematic manner, even in the absence of planning or of an explicit hypothesis-testing strategy, people are more likely to reach the correct conclusion.

The relationship between systematic exploration of the experiment space and successful performance has several implications. First, prior research has identified a connection between systematic explicit hypothesis-testing strategies such as VOTAT and successful performance (e.g. Vollmeyer et al., 1996). However, it is not clear that participants in these studies were explicitly forming and testing hypotheses when they conducted experiments. Possibly, their improved performance was related to an increasingly systematic search of the experiment space, rather than to a shift to an optimal explicit hypothesis-testing strategy. Clearly, such systematic search of the experiment space is also associated with people finding the correct solution.

Second, there are implications for instruction in scientific reasoning. It seems likely that more people are "experimenters" than "theorists" and therefore more likely to search the experiment space than the hypothesis space when presented with a scientific reasoning task. However, teaching explicit hypothesis-testing strategies *per se* may not lead to improved performance (e.g. Tweney, et al., 1980). It is possible that teaching strategies to conduct a systematic search of the experiment space, by using a systematic data collection method, might be an effective means of helping students improve performance on scientific reasoning tasks.

Acknowledgments

This research was supported in part by a student fellowship from George Mason University to the first author and by problem number 55-7294-A8 from the Office of Naval Research to the Naval Research Laboratory.

References

Baker, L. M., & Dunbar, K. (1996). Problem spaces in real-world science: What are they and how do scientists search them? *Proceedings of the 18th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: verbal reports as data.* Cambridge, MA: MIT Press.

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

Schauble, L., Klopfer, L. E., & Raghavan, D. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching, 28*(9), 859-882.

Schunn, C. D., & Klahr, D. (1995). A 4-space model of scientific discovery. *Proceedings of the 17th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Shute, V. J., & Glaser, R. (1990). A large-scale evaluation of an intelligent discover world: Smithtown. *Interactive Learning Environments, 1,* 51-77.

Shute, V. J., Glaser, R., & Raghavan, K. (1989). Inference and discovery in an exploratory laboratory. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences.* San Francisco: Freeman.

Siegler, R. S. & Atlas, M. (1976). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Detecting interactive patterns in data. *Journal of Educational Psychology, 68*(3), 360-370.

Simon, H. A., & Hayes, J. R. (1976). The understanding process: problem isomorphs. *Cognitive Psychology, 8,* 165-190.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development, 51,* 1-10.

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

Vollmeyer, R., Burns, B. D. & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20,* 75-100.

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