Changes in Learners' Exploratory Behavior in a Simulated Psychology Laboratory

Kazuhisa Miwa, Norio Ishii, Hitomi Saito, and Ryuichi Nakaike
{miwa, ishii, hitomi, nakaike}@cog.human.nagoya-u.ac.jp
Graduate School of Human Informatics, Nagoya University
Nagoya, 464-8601 JAPAN

Abstract

We constructed a virtual psychology laboratory (called VPL) on a computer. VPL simulates the process of pair subjects collaboratively solving Wason's 2-4-6 task, which has been traditionally used in the field of the psychology of discovery science. Participants were required to study collaborative problem solving while repeating experiments and hypothesis revisions using VPL. We conducted three experimental sessions using VPL. As a result, we confirmed, across the sessions, the improvement in various types of participant' performance, such as the organizational construction of experimental design, the degree of correctness of hypotheses the participants formed, and the generality of findings they discovered.

1. Introduction

It is one of the most important objectives in scientific research to understand the behavior of complex systems, such as physical, chemical, and biological systems. For example, psychologists, regarding humans as complex systems, try to identify the factors that determine the behavior of the systems (humans) through psychological experiments. Various types of knowledge are needed to organize psychological experiments. The ability to control experimental factors, CVS (the Control of Variables Strategy), is regarded as one of the most important skills. Klahr et al. have empirically studied the CVS ability of various types of subjects, such as elementary school students, university undergraduates, and graduates majoring in psychology, by analyzing the discovery process for programming grammar to manipulate a toy vehicle called BigTrak (Klahr, 2000). Moreover, they tried to apply the findings on CVS ability obtained in their laboratory studies to a real educational environment (Klahr, 2001).

Schunn and Anderson constructed a simulated psychology laboratory, called SPL, on a computer. Using SPL, they conducted an experiment in which university students and professional psychologists participated, and analyzed their abilities for designing and interpreting experiments (Schunn & Anderson, 1999). In their analysis, they discussed the difference between the general domain-independent and domain-dependent skills used by each participant for planning psychological experiments.

Additionally they proposed that SPL could be used as a learning environment for tutoring in experimental planning skills (Schunn & Anderson, 2001). However, in SPL, two ad hoc theories were given to the participants; and the participants were required to plan experiments that determined which of those two theories was valid. The process of forming theories (hypotheses) was ignored. Additionally, SPL did not actually simulate the human cognitive process, but simply output subjects' performances, using a previously installed function, as numeral values of the parameters input by the subjects. The process through which the output was obtained was not considered. In the present study, we construct a more realistic and complex experimental environment called VPL (Virtual Psychology Laboratory). Using VPL, we let university undergraduates experience conducting psychological experiments that lasted for several hours.

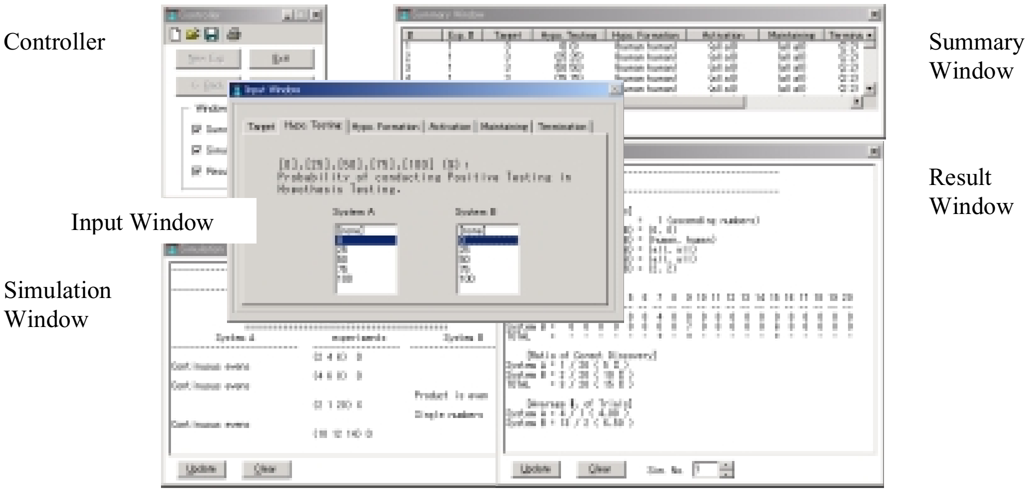
Schunn and Anderson were mainly interested in how the participants' behavior changed based on their degree of expertise in the research domain concerned. Our interest, on the other hand, is to show changes in the participants' behavior, such as in the formation and verification process of hypotheses including the stage of experimental planning, as a function of their training. We are also interested in the effect of VPL as a simulated psychology laboratory on the training and the improvement of learner's experimental behavior.

2. Experimental environment

2.1 VPL: Virtual Psychology Laboratory

In VPL, two production systems collaboratively solve a traditional discovery task: Wason's 2-4-6 task (Wason, 1960). The mission given to participants was to study factors determining the systems' performance. We can think of the factors determining the performance as, for instance, the degree of difference between the two systems' strategies, the interaction between those strategies and nature of targets, and the capacities of the systems' working memory.

It should be noted that this research theme being used for VPL is a highly realistic subject that has been



[0],[25],[50],[75],[100] (%):
Probability of conducting Positive Testing in Hypothesis Testing.

Figure 1: System’s interface.

discussed, through recent decades, by psychologists studying human collaborative discovery in laboratory studies (Gorman, 1992; Laughlin, et al., 1997). Moreover, it is also important to note that the psychological validity of this simulator has been tested by our several experiments. The system’s performance is determined according to an actual simulation of solving the task. We have already confirmed that the performance of this simulator reflects that of humans well (Miwa, 2001). Figure 1 shows the interface of VPL.

The "Controller" manages the starting and ending of simulations and the appearance of each window. The participants set up experimental factors in the "Input Window". The "Simulation Window" presents a real time process of two production systems solving the Wason’s 2-4-6 task. The "Result Window" shows the final result of each simulation. The "Summary Window" summarizes the experimental results obtained by the preceding simulations.

Table 1 shows the experimental factors that the participants can manipulate. Five of the six factors (excluding "Target") are specified in each of the two production systems. In the following experiment, the values of two parameters (# of activated instances and # of maintained hypotheses) were fixed at "all"; the participants could manipulate only the other four parameters. The performance of the simulator is determined by various factors. The fundamental nature of its behavior, such as the existence of interaction between the generality of the targets and the hypothesis-testing strategy (Klayman & Ha, 1987) and a main effect of the working memory capacity (# of activated instances and # of maintained hypotheses), is thoroughly consistent with the findings that several psychologists have reported in real psychological experiments.

2.2 Experiment

Participants: Twenty undergraduate students, not majoring in psychology, participated in the experiment as a part of a university class.

Background knowledge: Prior to the experiment, the participants learned the experimental procedure of Wason’s 2-4-6 task, and also the research objectives and motivations of laboratory studies using this kind of simple task. In a preliminary class, the participants read a research paper, which was experimental material prepared by the authors. The paper indicated the experimental result when a single subject solved the task. The result showed that there was interaction between the hypothesis-testing strategy and the nature

Table 1: Factors determining the simulator’s

Factors	Levels
Target	[#1] - [#35] Thirty-five kinds of targets used in the experiment. For example, Target #1 is "ascending numbers"; Target #35 is "three different numbers".
Hypothesis testing strategies	[0], [25], [50], [75], [100] The probability of conducting positive tests in generating instance [100] and [0] mean that the simulator always conducts positive tes and negative tests, respectively.
Hypothesis formation strategies	[human], [random], [specific], [general] [human] means that the simulator generates hypotheses as human do. [random]: generating hypotheses randomly. [specific]: generating specific hypotheses prior to general ones. [general]: generating general hypotheses prior to specific ones.
# of activated instances	[all], [6], [5], [4], [3] The number of instances that can be activated at once in the working memory when generating hypotheses.
# of maintained hypotheses	[all], [5], [4], [3], [2] The number of previously rejected hypotheses that can be maintained in the working memory.
Condition for terminating the search	[all], [5], [4], [3], [2] The number of continuos confirmations when the simulator terminates the search. [2] means when a hypothesis is continuous confirmed two times, the simulator recognizes the hypothesis as th final solution, and terminates the search.

of the targets. The participants took part in the experiment after understanding this finding.

Procedure: Three experimental sessions were conducted at intervals of a week. Each session lasted for one hour and a half. At the end of each session, the participants were required to report the findings they had obtained from a series of experiments in the session.

The participants' behavior in each experimental session basically repeated the following procedures. First, the participants entered, in the experimental sheet, (1) the objectives of the experiment they would perform (what are they investigating?), (2) the prediction of the experimental result, and (3) the experimental planning used for controlling experimental factors (which factors are focused on and which levels of each factor are searched?); then they performed the series of experiments planned in the experimental sheet, by manipulating the simulator. After obtaining the experimental result, they entered (4) the interpretation of the experimental result. The participants repeated this series of procedures until the end of the session.

Pre- and Post- tests: Before and after the three experimental sessions, pre- and post- tests were conducted to measure the subjects' fundamental ability to control experimental factors.

3. Experimental results

3.1 Chunking behavior

We define a set of organized experiments as a chunk. Thus, we think of a more sophisticated construction of experimental planning as a process of constructing higher chunks (Miwa, 2000).

The participants conducted their experiments by searching the experimental space as depicted in Figure 2. As mentioned before, two factors, # of activated instances and # of maintained hypotheses, were fixed at the value "all". The participants manipulated the simulator and obtained experimental results after filling in the experimental sheet. A set of experiments planned in a piece of the experimental sheet is regarded as a unit of experiments. Almost all experimental planning (about 96%) entered in a piece of the experimental sheet was constructed based on the factorial experiment design. So, for example, when p levels and q levels in each of two factors were searched, a total of p x q experiments was completely performed according to the experimental planning. We excluded, from the following analysis, units of experiments (4%) which violated this factorial experiment design.

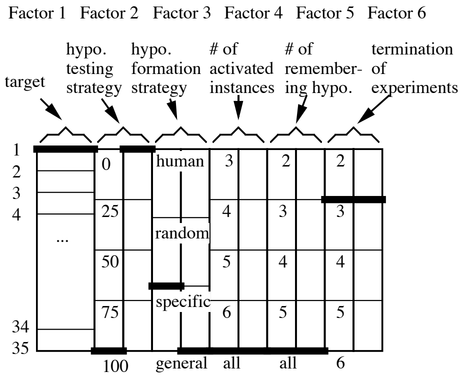
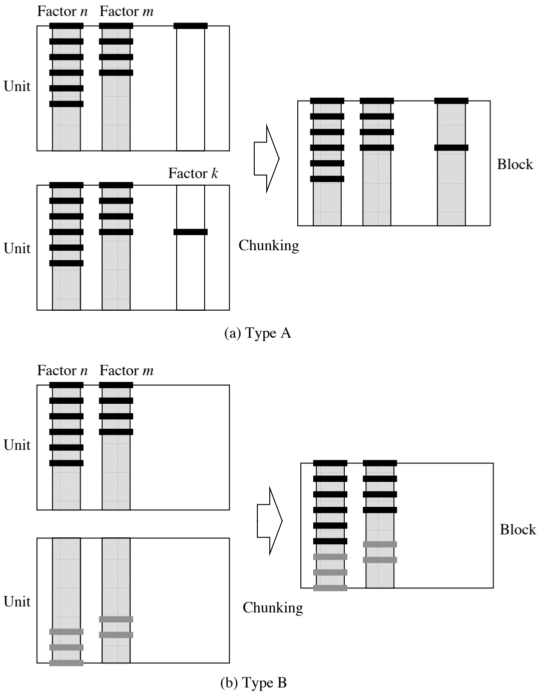


Figure 2: The experimental space searched by Ss.

We regard this set of experiments planned in a piece of the experimental sheet as the most basic chunk. We call this basic chunk a "Unit". The participants combine multiple Units to construct a higher chunk. We propose the following two types of chunking, Type A and Type B, as methods for constructing a higher chunk.

See Figure 3 in which a Unit is constructed by the set of experiments where some levels of Factor *n* and Factor *m* are searched. The first type of chunking is Type A (Figure 3(a)) where the searcher shifts a searching level of Factor *k* one by one, while maintaining the search of Factor *n* and Factor *m*. The set of these experiments can be grouped as a chunk of experiments in which three factors, Factor *n*, Factor *m*, and Factor *k*, are simultaneously controlled. The important point is that factors other than the controlled



(a) Type A

(b) Type B

Figure 3: Two types of chunking behavior.

three factors are fixed at an identical level.

The second type of chunking, Type B (Figure 3(b)), occurs when it is impossible that all levels involved in the focused factors, such as Factor *n* and Factor *m*, can be searched at the same time; the search is divided into multiple Units. In this case, the set of multiple Units can also be seen as a chunk. The point is that factors other than Factor m and n are fixed.

By using these two types of chunking, bigger chunks can be constructed from multiple basic Units. We call these higher chunks "Blocks". Here we define the compression ratio of chunking based on the number of individual experiments constructing a single Block. For example, in Figure 3 (a), one Block is constructed from 48 experiments [= 6 (Factor n) x 4 (Factor m) x 2 (Factor k)]; so the compression ratio of chunking is 0.021 (= 1/48). On the other hand, in Figure 3(b), as 30 experiments (= 6 x 4 + 3 x 2) construct a Block, the compression ratio of chunking is 0.033 (= 1/30). The smaller ratio of chunking means that the participants are able to construct a bigger chunk in their experimental behavior. Consequently, the compression ratio of chunking reflects the degree of participants' organizational experimental behavior.

Figure 4 shows, for each of the three experimental sessions, the average compression ratio of chunking of the 16 out of 20 participants, who participated in all of the three experimental sessions. (Similarly analyses of these 16 subjects' results are shown in sections 3.2, 3.3, and 3.4.) As the experimental sessions proceeded, the compression ratio of chunking decreased. As a result of ANOVA, a main effect of the experimental sessions was significance ($p < .01$). This result confirms that the participants learned to construct bigger chunks, i.e., exhibited more organizational behavior, through repeating experimental activities.

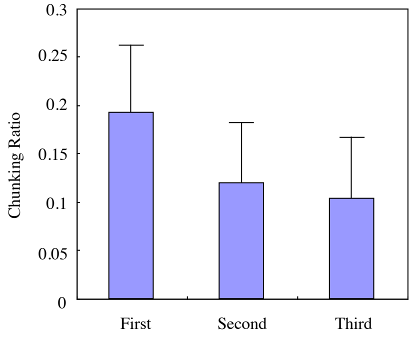


Figure 4: Transition of the ratio of chunking.

3.2 Controlled factors

We can also confirm the process of constructing bigger chunks by analyzing the transition of the number of controlled factors by the participants across the three experimental sessions. Figure 5 shows the average ratios of the number of Blocks, in which one, two, and three or more factors were controlled, to the number of all Blocks. As the experimental sessions proceeded, the ratio of Blocks manipulating more than three factors increased, whereas the ratio of Blocks manipulating one factor decreased. As a result of ANOVA, there was interaction between the experimental sessions and the number of controlled factors ($p < .05$). A simple main effect of the experimental sessions at each of the two single levels, one and more than three, in the number of controlled factors was significance ($p < .05$ and $p < .01$ respectively). The result above shows that the participants learned, during the progress of the experimental sessions, to conduct experiments in which a greater number of various factors were manipulated.

3.3 Hypotheses

We also focused on the hypotheses formed by the participants.

The participants entered their prediction of the experimental results in the experimental sheet before executing a series of experiments. At that time, they also estimated the degree of confidence in the prediction on a 1 to 5 scale. Additionally, after executing the experiments with the simulator, they entered their interpretation of the experimental results. At that time, they also estimated the degree of correctness of their prediction on a 1 to 5 scale.

Figure 6 shows the average degree of confidence estimated before executing experiments and the average degree of correctness estimated after the experiments. The figure indicates that the degree of correctness was improved from the first to third sessions while the degree of confidence was almost constant. As a result of ANOVA, there was interaction between the experimental sessions and the two kinds of participants' estimation (the degree of confidence and correctness) ($p < 0.01$). A simple main effect of the experimental

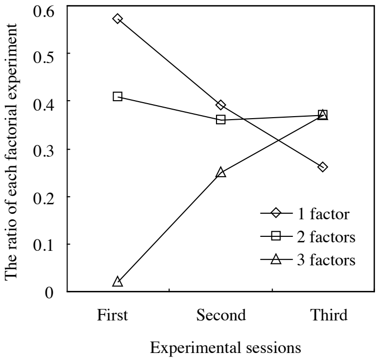


Figure 5: Transition of the number of controlled factors.

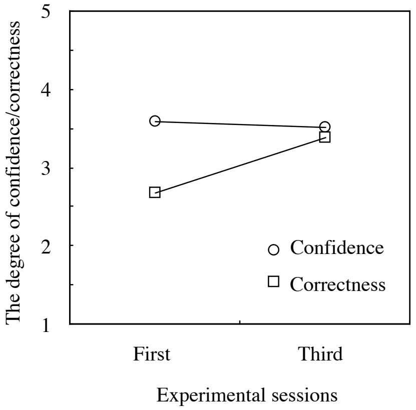


Figure 6: Transition of the degree of confidence/correctness of hypotheses.

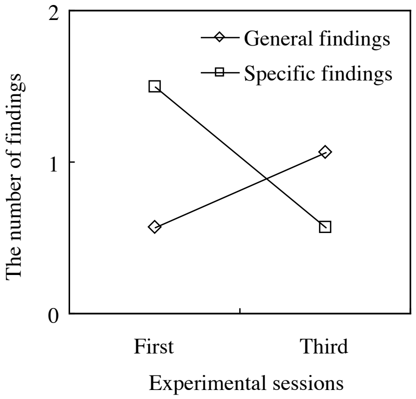


Figure 7: Transition of the number of specific/general findings.

sessions at the degree of correctness revealed significance ($p < 0.05$) whereas an effect at the degree of confidence did not.

The degree of correctness reflects the objective validity of the participants' hypotheses whereas the degree of confidence reflects the participants' subjective estimation of the probability of their hypotheses. The invariant of the degree of confidence implies that the change in the complexity of the participants' hypotheses was not so marked between the former and latter parts of the experimental sessions. On the other hand, the improvement in the degree of correctness confirms that the participants learned to form more accurate hypotheses during the progress of their experiments even though the complexity of the hypotheses was almost constant.

3.4 Findings

Next we move to an analysis of the findings that the participants discovered. As mentioned before, the participants were required to report their findings at the end of each experimental session.

We categorize the findings from the viewpoint of their generality. We define participants' general conclusions mentioning the relation between an experimental factor (or factors) and the system's performance as general findings. For example, the conclusion, "positive testing is effective in finding the specific targets whereas negative testing is effective in finding the general targets", is an example of a general finding because the participants mention the relation between the two factors, the nature of targets and the hypothesis-testing strategies, and the system's performance. On the other hand, we define restricted conclusions mentioning a factor (or factors) determining the system's performance only in a specific situation as specific findings. For example, the conclusion, "in terms of target #27, negative testing is effective ", is an example of a specific finding because this conclusion mentions a restricted finding for a specific case: target #27.

Figure 7 shows the average number of specific and general different findings in the first and third experimental sessions. The figure shows that the number of general findings increased across the sessions while the number of specific findings decreased. As a result of ANOVA, there was interaction between the experimental sessions and the nature of findings (specific and general) ($p < 0.01$). Simple main effects of the experimental sessions at both levels of specific and general in the nature of findings revealed significance ($p < .01$ and $p < .05$ respectively). This confirms that the participants gradually came to discover general findings during the progress of the experimental sessions.

3.5 Improvement from Pre test to Post test

Lastly, we discuss whether the participants learned general procedural knowledge on experimental planning by analyzing the pre- and post- tests that were carried out before and after all of the experimental sessions.

In the pretest, the participants were required to plan an experiment that identified the factors (temperature and/or humidity) responsible for the growth of bacteria. In the posttest, an isomorph of the problem in the pretest was used where the participants were required to identify the factors causing the growth of plankton. The participants' solutions in each test were categorized into two types: (1) for identifying the factors determining the growth of bacteria or plankton, first varying one factor while fixing the other factor then manipulating that other factor (that is, first varying humidity while fixing temperature then varying temperature while fixing humidity); and (2) simultaneously controlling both two factors. We call

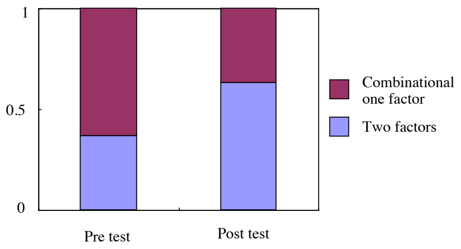


Figure 8: The comparison of experimental planning in the pretest and the posttest.

the former planning a "combined one factor experiment", and the latter a "two factors experiment". The latter planning is more sophisticated because it can detect interaction between the two factors, but the former cannot. Figure 8 shows the comparison of the solutions of 19 participants in the pretest and in the posttest. One of the 20 participants was excluded from the analysis because the subject indicated a confusing answer. Fisher's exact analysis supported a tendency in the increase of the two factors experiment in the posttest compared to in the pretest ($p < .1$).

The above result confirms that some of the participants successfully acquired general procedural knowledge on conducting appropriate experimental planning through repeatedly performing experiments using VPL.

4. Discussions and conclusions

In this experiment, the participants were not given any instruction from a tutor. The participants experienced the three experimental sessions receiving the feedback from the simulator while repeatedly performing their experiments by themselves without any instruction from others. However, the various types of participants' performance, such as organizational designing of experiments, the degree of correctness of formed hypotheses, and the generality of findings, were remarkably improved. This implies that this kind of exercise using a simulated research environment, such as VPL, could be effective for providing tutoring in psychological activities to students who begin to learn experimental psychology.

We understand that it was still not clear that these improvements were brought about by the learning of general experimental skills such as CVS or simply by the increase of information on the problem space searched during the progress of the experiments. However, we believe that the improvement of the scores from the pretest to the posttest confirms that some of the participants had learned something related to general skills on experimental planning because the contents of those tests were independent from those dealt with in the exercise using VPL. At any rate, the experimental results support the possibility of achieving "learning by doing" without instructions through this sort of relatively short-term exercise by using a VPL-like learning environment (Anzai, 1979).

In our future work, we will examine the usage of VPL as an experimental microworld. We could clarify, for instance, the difference between Novices' and Experts' experimental processes and the effects of background knowledge on the processes. We will also further discuss on the possibility of using VPL as a tutoring system. For example, it might be possible to activate the participants' learning process by giving informative feedback to learners based on the idea of constructing higher chunks.

References

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. Psychological Review, 86, 124-140.

Gorman, M. (1992). Simulating science: heuristics, mental models, and technoscientific thinking. Indiana university press.

Klahr, D. (2000). Exploring science: The cognition and development of discovery processes. Cambridge, Mass.: MIT Press.

Klahr, D., Chen, Z., & Toth, E. (2001). From cognition to instruction to cognition: a case study in elementary school science instruction. In C. Crowley, et al. (Eds.), Designing for science: implications from everyday, classroom, and professional settings. Mahwah, NJ: LEA.

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

Laughlin, P. R., Magley, V. J., & Shupe, E. I. (1997). Positive and negative hypothesis testing by cooperative groups, Organizational behavior and human decision processes, 69, 265-275.

Miwa, K. (2000). Human Discovery Processes Based on Searching Experiments in Virtual Psychological Research Environment. LNAI, 1967, 225-239.

Miwa, K. (2001). Emergence of effects of collaboration in a simple discovery task. Proceedings of the 23rd annual conference of the cognitive science society, 645-650.

Schunn, C. D., & Anderson, J. R. (2001). Acquiring expertise in science: explorations of What When and How. In C. Crowley, et al. (Eds.), Designing for science: implications from everyday, classroom, and professional settings. Mahwah, NJ: LEA.

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. Cognitive Science, 23, 337-370.

Wason, P. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly journal of experimental psychology, 12, 129-140.