Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE @ DIRECT®

Cognitive Systems Research 5 (2004) 41–62

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Collaborative discovery in a simple reasoning task

Action editor: Lynne Reder

Kazuhisa Miwa

Graduate School of Information Science, Nagoya University, Furo-cho, Nagoya-shi 464-8601, Japan

Received 10 February 2003; accepted 18 July 2003

Abstract

In psychological experimental studies on scientific discovery, discussions have been made on the effects of collaborative discovery using simple experimental tasks, such as Wason's 2-4-6 task. Generally speaking, however, these studies have not explained the prominent effects of multiple subjects collaboratively discovering a target. In the present study, we identify situations in which the effects of collaboration emerge, and why they emerge, by combining a computer simulation method that employs our computational model as a cognitive simulator with a method based on laboratory studies. We basically control two factors, i.e. the hypothesis-testing strategy used by the subjects and the nature of the target that the subjects are required to find, both of which have been identified by previous psychological studies as key factors determining the subjects' performance. The computer simulations show that the performance of combined two systems in collaborative discovery exceeds that of each system in independent discovery, but only when the two systems try to find a target having the nature of generality by repeatedly conducting a positive test. This finding is also confirmed by psychological experiments designed to verify the computer simulations. Moreover, through a theoretical analysis, we show that this effect of collaboration is provided by the emergence of negative tests from the interaction of the two systems (or humans) repeating only a positive test.

Keywords: Collaboration; Problem solving; Scientific discovery; Hypothesis testing

1. Introduction

Many laboratory studies on scientific discovery have so far used relatively simple tasks, such as Wason's 2-4-6 task and New Elusis (Gorman, 1992; Newstead & Evans, 1995). Recently, by using such tasks, the effects of collaboration have been empirically discussed in the case of several participants collaboratively finding a target.

Many arguments have been given for the benefits of collaboration and science. In this research, we focus on the effects of collaboration in a domain-independent and knowledge-free hypothesis discovery situation, which can be captured using the psychology laboratory method. One of approaches is to simplify both the nature of the task and the type of the collaboration, and Wason's

E-mail address: miwa@is.nagoya-u.ac.jp (K. Miwa).

1389-0417/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2003.10.001

task is one such well-studied, highly simplified example. Some important aspects of scientific discovery may be missed through this type of approach, but I believe that there are important characteristics that can be clarified by proceeding in this way (see detailed discussions in Chapter 1 of Klahr, 2000).

It is commonly believed that people working together provides positive effects: e.g. reducing the possibility of making mistakes through mutual testing, finding another representation of problems by obtaining a different viewpoint from another person, and activating the generation of new innovative ideas through brainstorming. However, some of the empirical results obtained in the above psychological studies have not consistently supported this intuitive prediction.

In the present study, to estimate the benefits of collaboratively finding a target, we introduce three kinds of conditions: a single condition, an independent condition, and a collaborative condition, as shown in Fig. 1. In the single condition, a solo subject discovers a target while forming his/her own hypothesis and experiments. In the independent condition, two subjects discover a target but no interaction is permitted between them. In the collaborative condition, two subjects interactively discover a target through conversation, i.e. exchanging their hypotheses and sharing their experimental results.

In laboratory studies, the performance (proportion of correct findings) in the single condition (in which a single subject performs the task) and that in the collaborative condition (in which a group of n subjects collaboratively performs the task) are compared. If anybody in the group discovers the solution then the group is scored as having discovered the solution. In this comparison, it is shown that even when the latter performance exceeds the former, the advantage may be provided not by the interaction among the subject, but simply by the number of the subjects. That is, in the latter case of n solutions (final hypotheses) by n subjects, the probability that at least one of the solutions is identical to the target is greater. Accordingly, we also consider the independent condition, in which n participants independently perform the same task without interaction. The performance in the independent condition is then theoretically calculated from the performance in the single condition. That is, the probability that at least one of the n subjects reaches the solution is $1 - (1 - p)^n$, where the probability of each subject finding the correct target is p $(0 < p < 1)$. We utilize this score as the performance evaluation in the independent condition.

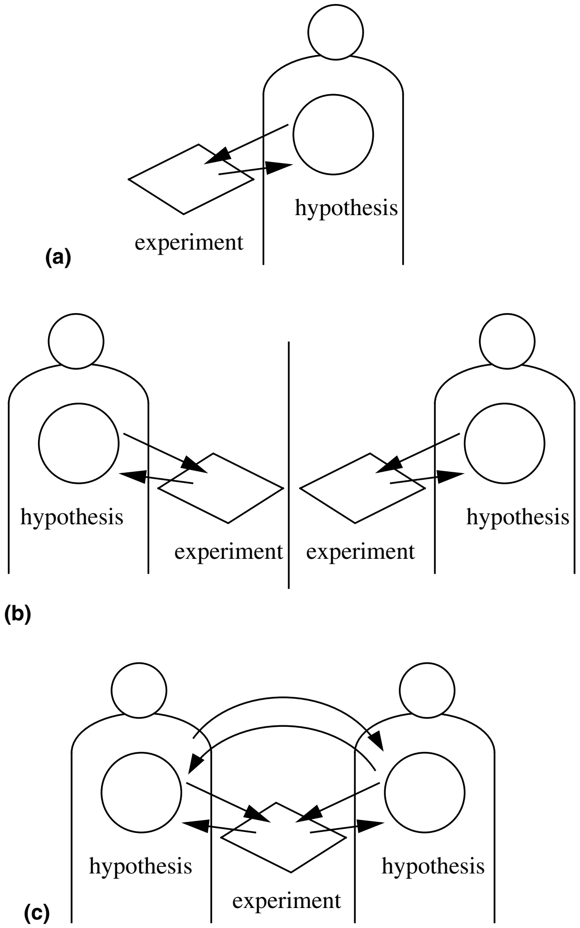


Fig. 1. Three stages of interaction. (a) Single condition. (b) Independent condition. (c) Collaborative condition.

Table 1 compares performances among the single, independent, and collaborative conditions in previous studies that used simple reasoning tasks (Freedman, 1992; Laughlin & Futoran, 1985; Laughlin & McGlynn, 1986; Laughlin et al., 1991; Laughlin et al., 1998). The performances of the independent condition of Laughlin and Futoran (1985), Laughlin and McGlynn (1986) and Freed-

Table 1
Comparison of performances in the three stages of interaction in the previous studies

	Laughlin and Futoran (1985)	Laughlin and McGlynn (1986)	Laughlin, VanderStoep, and Hollingshead (1991)					Laughlin, Bonner, and Altermatt (1998)	Freedman (1992)	
# of group members	4	4	4					4	4	
Task	New Elusis	New Elusis	New Elusis					New Elusis	2-4-6 Task	
Single	0.15	0.19	0.06	0.13	0.15	0.14	0.16	0.16	0.33	0.08
Independent	0.47	0.57	0.2	0.3	0.38	0.3	0.41	0.38	0.80	0.28
Collaborative	0.35	0.34	0	0.1	0.2	0.1	0.41	0.34	0.83	0.67

man (1992) were theoretically calculated by the procedure above mentioned. On the other hand, the performances in the independent condition of Laughlin et al. (1991, 1998) were actual empirical data. In the two papers, the independent condition in which four subjects independently (without interaction) solved the task was actually set up. The performances of the best, 2nd-best, 3rd-best, and 4th-best participants were indicated. Therefore, in the table the performances of the best participants are used as those in the independent condition and the mean scores of the four participants as the performances in the single condition.

The table shows that the performance in the collaborative condition fails to exceed that in the independent condition in almost all cases. However, the analyses so far have been relatively coarse grained and do not address the dynamic or the mechanisms of instance generation or hypothesis formation. This negative tendency of the effect of collaboration is also seen in the studies of idea generation. For instance, Paulus and Huei-Chuan presented much research demonstrating that idea sharing in groups involves relatively inefficient processes (Paulus, 2000; Paulus & Huei-Chuan, 2000).

As explained later, there are some factors that strongly influence the performance of subjects discovering a target. In the above experiments, however, these important factors are not necessarily controlled. The reasons for this experimental inadequacy include several practical factors, such as the cost for performing the experiments, as discussed in Section 2.

Consequently, there is need for further investigation. The main goals of this study are to identify the conditions or features of rule discovery tasks in which collaboration is or is not more successful than a simple aggregation of independent problem solvers. Note here that the current paper only discusses dyadic scientific discovery, i.e. collaboration based on two-person groups.

There are various methodologies that can be used in studies on scientific discovery. Each has its own methodological advantages, so some have stressed the importance of combining different approaches (Klahr, 2000; Klahr & Simon, 1999). In this study, we attempt to overcome by improving the above-mentioned disadvantages by unifying several research methodologies. We first propose a hypothesis on when the effects of collaboration appear by using a computational model that solves Wason's 2-4-6 task (Wason, 1960) in computer simulations. We then verify the hypothesis by psychological experiments. Lastly, we generalize the empirical findings by theoretical task analysis and discuss why the effects emerge only in specific situations.

2. Approaches

The laboratory studies mentioned above reveal several weak points as methodologies for research.

First, in experiments in which various experimental factors are controlled, it is necessary to have many subjects participate in the experiments. Particularly in collaboration studies, a greater number of controlled factors significantly increases the required number of participants. We have to cross exponentially increasing combinations of subjects with different conditions. Generally speaking, when we want to control n conditions, $(n \times (n+1))/2$ sets of experiments are needed in the collaborative problem solving case, whereas only n sets are needed in the single problem solving case. Naturally, the increase in the cost for performing the experiments often creates practical difficulties in the research.

Second, the control of experimental factors is usually carried out by an experimenter providing instructions to participants. For example, an experimenter may instruct a participant in an experimental setting by saying: "Please perform the hypothesis testing only by using a positive test", or "You must always form two hypotheses (or only one hypothesis) at once throughout the experiments". Of course, some subjects sometimes may not follow the instructions. When we analyze experimental results, we usually find that we have to exclude irregular data by subjects who violated instructions. However, in studies on collaboration, if at least one subject comprising a collaborative group violates a procedure, then none of the data of that group involving the problematic subject can be used. Again, we face the difficulty of increasing the experimental cost in the research.

To overcome the above difficulties, we have used a computational model as a cognitive simulator (Miwa, 1999, 2001; Miwa, Ishii, Saito, & Nakaike, 2002). The simulator actually runs on a computer, and its behavior can be controlled by manipulating the parameters of the simulator.

If we were able to construct a cognitive simulator that could properly demonstrate the human problem-solving process, then we could perhaps solve the methodological difficulties in laboratory studies by simulating discovery processes in various situations on a computer. That is, if a computational model could generate a large set of predicted outcomes, then an investigator could select, prior to performing actual experiments, the important experimental settings from among them that would give the most interesting, non-obvious, and counter-intuitive results. Therefore, this approach might resolve the efficiency issue discussed earlier with respect to the costly use of human subjects.

3. Background

In this study, we use Wason's 2-4-6 task as an experimental task. The standard procedure of the 2-4-6 task is as follows. All subjects are required to find a rule of a relationship among three numerals. In the most popular situation, a set of three numerals, "2, 4, 6", is presented to subjects at the initial stage. The subjects form a hypothesis about the regularity of the numerals based on the presented set. The subjects then produce a new set of three numerals and present it to the experimenter. This set is called an instance. The experimenter gives a Yes as feedback to the subjects if the set produced by the subjects is an instance of the target rule, or a No as feedback if it is not an instance of the target rule. The subjects continuously carry out experiments, receive feedback from each experiment, and search to find the target. The subjects propose a final hypothesis whenever they think they know what the rule is and they receive feedback on whether or not the hypothesis is correct.

Klayman and Ha (1987) gave some decisive answers to several important questions that had been discussed in psychological studies using traditional discovery tasks such as Wason's task

(Klayman & Ha, 1987, 1989). One of their major conclusions was that there is substantial interaction between the nature of discovered targets and the effectiveness of hypothesis-testing strategies used by subjects. In this study, we mainly control these two factors in our experiments.

First, we briefly explain important concepts regarding the two key factors, i.e. the nature of the targets that the subjects try to find and the hypothesis-testing employed by the subjects.

3.1. The nature of targets

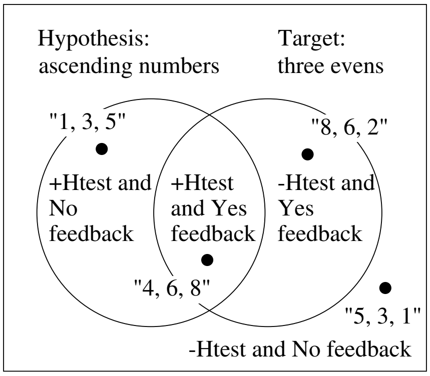
We categorize the targets used in our experiments from the viewpoint of their generality. We define targets as broad targets if the proportion of their members (positive instances) to all instances (all sets of three numerals) in the search space is large. On the other hand, we define targets as narrow targets if the same proportion is small. An example of the former type of target is "the product of three numerals is even" (where the proportions of target instances to all possible instances is 7/8) and an example of the latter type is "three evens" (where the proportion is 1/8).

3.2. Hypothesis testing

There are two types of hypothesis testing: a positive test and a negative test. The positive test (+Htest) is conducted in an instance where the subject expects there to be a target. That is, the +Htest is a hypothesis test using a positive instance for a hypothesis. The negative test (−Htest) is, in contrast, a hypothesis test using a negative instance for a hypothesis. For example, if a hypothesis were about "ascending numbers", the +Htest would use a sequence like "1, 3, 9"; the −Htest would use a sequence like "1, 5, 2". There are many types of hypotheses-testing strategies and they can be defined in terms of the extent to which they exclusively use either positive tests or negative tests. In the following, we use the term "+Htest strategy" when we use only +Htests throughout experiments.

In the following description, to avoid confusing basic concepts, we define Yes and No instances as members and non-members for targets, which the subjects do not know. On the other hand, +Htests and −Htests are defined by whether instances tested are members or non-members for hypotheses, which the subjects form. The subjects can discriminate whether a certain set of three numerals is a positive or negative instance for their hypotheses but cannot know whether it is a Yes or No instance until the experimental feedback is given. When a subject conducts an experiment using a positive instance for his/her hypothesis and knows, through the feedback from the experimenter, that the instance is a Yes instance for a target, we say that the subject receives a Yes feedback as a result of his/her +Htest.

Klayman and Ha summarized states in which a subject's hypothesis is falsified. Fig. 2 illustrates these states in the example situation where the target is "three evens" and the subject's hypothesis



Hypothesis	Target: Yes	Target: No
+Htest	ambiguous verification	conclusive falsification
-Htest	conclusive falsification	ambiguous verification

Fig. 2. Patterns of confirmation vs. disconfirmation.

is "ascending numbers". When the subject conducts a +Htest using the instance "1, 3, 5" and then receives a No feedback, his/her hypothesis is disconfirmed (false positives). Another state of conclusive falsification is caused by the combination of a −Htest and a Yes feedback, using the instance "8, 6, 2" (negative hits). On the other hand, states of ambiguous verification are obtained from the combination of a +Htest and a Yes feedback, using "4, 6, 8" (positive hits), or the combination of a −Htest and a No feedback, using "5, 3, 1" (false negatives). The importance of the function of falsification has been stressed from the viewpoint of the philosophy of science (Popper, 1959).

4. Computer simulations

In this section, we perform computer simulations for hypothesizing situations in which the effect of collaborative discovery emerges.

4.1. Example behavior of the model

First, to help the reader easily understand the basic specifications of this simulator, an example behavior of our model is shown in Table 2. In this case, two systems tried to find a target, i.e. each number is a divisor of 12, collaboratively. One system, System A, always used a +Htest in its experiments, and the other, System B, used a −Htest. Human participants sometimes ignored instances that had been observed so far when forming a hypothesis. However, the computational model always formed a complete hypothesis that was consistent with all instances that had been observed so far.

The experiments were alternately conducted. Through each simulation, one system generated half of all instances, and the other generated the other half. Instances are basically generated randomly by the following procedure. First, a candidate of an instance, each numeral of which is an integer ranging from −20 to 20, is generated randomly, then if the set of integer fits with the condition given, the candidate is approved of as an instance. Each experimental result was shared by both systems, that is, each system knew all generated instances with the Yes or No feedback given to each instance.

The left-most and right-most columns of the table indicate hypotheses formed by System A and System B, respectively. The middle column indicates experiments, that is, generated instances, Yes or No feedback, and whether a +Htest or −Htest was conducted and by which system. The left-most number in each column, from #1 through #41, indicates a series in the processing.

The distinction of ambiguous verification and conclusive falsification is determined by Klayman and Ha's normative schema as indicated in Section 3. In the experiments, System A disconfirmed its hypotheses at #4, #10, and #16, which were caused by self-conducted (i.e. conducted by System A) experiments at #3, #9, and #15. System B disconfirmed its hypotheses at #17 and #29, which were caused by other-conducted (i.e. conducted by System A) experiments at #15 and #27.

4.2. Model

The model was constructed on an interactive production system architecture that we had developed for simulating collaborative problem solving processes. Fig. 3 shows an outline of the architecture. The architecture primarily consists of five parts: production sets of System A, production sets of System B, working memory of System A, working memory of System B, and a common shared blackboard. The two systems interact through the common blackboard. That is, each system writes elements of its working memory on the blackboard and the other system can read them from the blackboard.

Our model organizes the knowledge of identifying the regularity of numerals in the form of a hypothesis space. Table 3 shows a hypothesis space consisting of 11 dimensions and the sets of their values. The model searches this hypothesis space and chooses one of the hypotheses that are consistent with all instances with the Yes or No labels recorded in the memory at that point. Basically, the model searches the hypothesis space randomly in order to generate a hypothesis. However, there are three particular hypotheses, i.e.

Table 2
Example behavior of the simulator

Hypothesis of System A		Experiments				Hypothesis of System B	
		1	2, 4, 6	Yes			
2	Continuous evens						
		3	4, 6, 8	No	+Htest by System A		
4	The product is 48						
						5	The sum is a multiple of 4
		6	6, 6, −17	No	−Htest by System B		
						7	The sum is a multiple of 4
8	The product is 48						
		9	24, −1, −2	No	+Htest by System A		
10	First + second = third						
						11	The sum is a multiple of 4
		12	3, −8, −20	No	−Htest by System B		
						13	The sum is a multiple of 4
14	First + second = third						
		15	−10, 2, −8	No	+Htest by System A		
16	Divisors of 12						
						17	The second is 4
		18	−5, −14, −9	No	−Htest by System B		
						19	The second is 4
20	Divisors of 12						
		21	4, 4, 8	Yes	+Htest by System A		
22	Divisors of 12						
						23	The second is 4
		24	−17, 3, 12	No	−Htest by System B		
						25	The second is 4
26	Divisors of 12						
		27	2, 12, −12	Yes	+Htest by System A		
28	Divisors of 12						
						29	Divisors of 12
		30	8, 12, −2	No	−Htest by System B		
						31	Divisors of 12
32	Divisors of 12						
		33	2, 6, −2	Yes	+Htest by System A		
34	Divisors of 12						
						35	Divisors of 12
		36	−2, −7, −8	No	−Htest by System B		
						37	Divisors of 12
38	Divisors of 12						
		39	4, 3, −12	Yes	+Htest by System A		
40	Divisors of 12						
						41	Divisors of 12

"three continuous evens", "the interval is 2", and "three evens". Human subjects tend to generate these hypotheses at first when the initial instance of "2, 4, 6" is presented. Therefore, the model also generates these hypotheses prior to the other possible hypotheses.

The model used in this simulation searched its hypothesis space everywhere any target was involved without exception, and it generated a hypothesis consistent with the instances that had been observed. In this sense, the model was regarded as solving a search task, whereas the real Wason's task is psychologically characterized as a discovery task.

In addition, the search strategy used in our model was too simplified from the viewpoint of psychological reality. For example, in this study, the degree of generality was only dealt with as a nature of

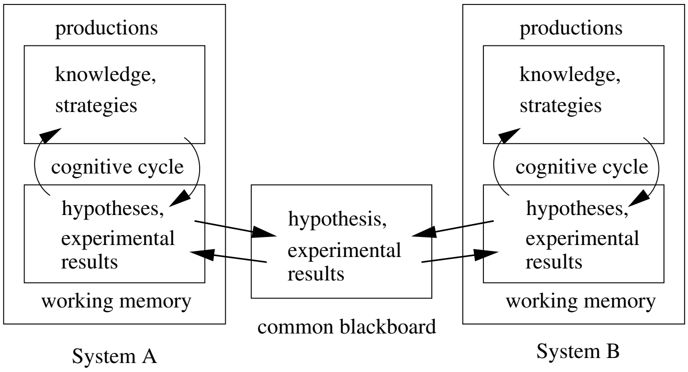


Fig. 3. Interactive production system architecture.

Table 3
Hypothesis space of the simulator

Dimensions	Values
Order	Ascending, descending, same digits, equal or ascending, equal or descending, up and down, down and up
Interval	Special interval *n*, same interval, increasing interval, decreasing interval
Even–odd	Continuous evens, continuous odds, three evens, three odds, even–even–odd, even–odd–even, odd–even–even, odd–odd–even, odd–even–odd, even–odd–odd
Range of digits	Single digits, double digits, positive digits, negative digits
Certain digit	even-?-?, ?-even-?, ?-?-even, odd-?-?, ?-odd-?, ?-?-odd, a certain digit *n* in an *m*th slot
Multiples	Multiples of *n*
Divisors	Divisors of *n*
Sum	Even, odd, single, double, positive, negative, certain number, multiple of *n*
Product	Even, odd, single, double, positive, negative, certain number
Different	Three different numbers
Mathematical relationship	First + second = third, first + third = second, second + third = first, first × second = third, first × third = second, second × third = first, second = 2 × first and third = 3 × first, second and third are multiples of first, third = first × second − 2

targets. However, there are another important factors such as the degree of familiarity. For example, two hypotheses, divisors of 24 and the first numeral is 2, both of which can be inferred from an initial instance “2, 4, 6”, have almost the same generality (see Table 4). However, the former target is likely to be more familiar than the latter. However, familiarity properties are not addressed in this study.

The way of searching the hypothesis space is strongly related to special matters of Wason’s task. However, what was intended to be dealt with in this paper is to measure the effects of collaboration based on a syntactic relation between the nature of targets (the generality of targets in this case) and a participant’s strategy (the hypothesis-testing strategy). Moreover, in actual psychological experiments, a search strategy seems very different with each individual. So we thought that it would be more important to simulate a situation in which two computational agents that use the same search strategy solve the task collaboratively, rather than tune up our model for accurately tracing the specific strategy of hypothesis space search used by each individual.

4.3. Design

In the computer simulations, we had the two systems find the 35 kinds of targets shown in Table 4. The proportion of each target instances to all instances was calculated, where the set of all instances is a set of three integers each of which ranges from −20 to 20 (so the number of all instances is 68921 (= 41 × 41 × 41)). The informa-

Table 4
Targets used in the simulations

No.	Rules	Proportions of instances	Nature
#1	Ascending numbers	15.5	Broad
#2	Equal or ascending numbers	17.8	Broad
#3	The interval is 2	0.1	Narrow
#4	The interval is the same	1.2	Narrow
#5	Continuous evens	0.05	Narrow
#6	Three evens	13.4	Broad
#7	Single digits	10.0	Narrow
#8	Positive digits	11.6	Narrow
#9	The first number is even	51.2	Broad
#10	The second is even	51.2	Broad
#11	The third is even	51.2	Broad
#12	The first is 2	2.4	Narrow
#13	The second is 4	2.4	Narrow
#14	The third is 6	2.4	Narrow
#15	Multiples of 2	13.4	Broad
#16	Divisors of 12	2.1	Narrow
#17	Divisors of 24	2.6	Narrow
#18	The sum is even	50.0	Broad
#19	The sum is a double digit	66.1	Broad
#20	The sum is a positive number	49.1	Broad
#21	The sum is 12	1.6	Narrow
#22	The sum is a multiple of 12	8.3	Narrow
#23	The sum is a multiple of 6	16.7	Broad
#24	The sum is a multiple of 4	25.0	Broad
#25	The sum is a multiple of 3	33.3	Broad
#26	The sum is a multiple of 2	50.0	Broad
#27	The product is even	88.4	Broad
#28	The product is a double digit	10.1	Narrow
#29	The product is a positive number	46.4	Broad
#30	The product is 48	0.2	Narrow
#31	The third = the first + the second	1.8	Narrow
#32	The third = the first × the second − 2	0.5	Narrow
#33	The third = the first × 3 and the second = the first × 2	0.02	Narrow
#34	The third = the first × n and the second = the first × m	7.9	Narrow
#35	Different three numbers	92.8	Broad

tion on the ratios was added to Table 4. Based on this information, the narrow and broad rules were defined concretely.

The initial instance was “2, 4, 6”. For each target, we executed 30 simulations to calculate the percentage of correct solutions.

The computer simulations were basically conducted based on the following 2 × 3 experimental design.

4.3.1. The nature of targets

We divided the 35 targets into two categories: (a) 18 narrow targets and (b) 17 broad targets.

4.3.2. Hypothesis-testing strategies

Three combinations of hypothesis-testing strategies were investigated. They were (a) +Htest and +Htest, (b) −Htest and −Htest, and (c) +Htest and −Htest.

4.4. Results

Fig. 4 shows the results of the computer simulations. The horizontal axis of each figure indicates the number of experiments, that is, the number of generated instances, whereas the vertical axis indicates the proportion of correctly finding the 18 narrow targets and the 17 broad targets.

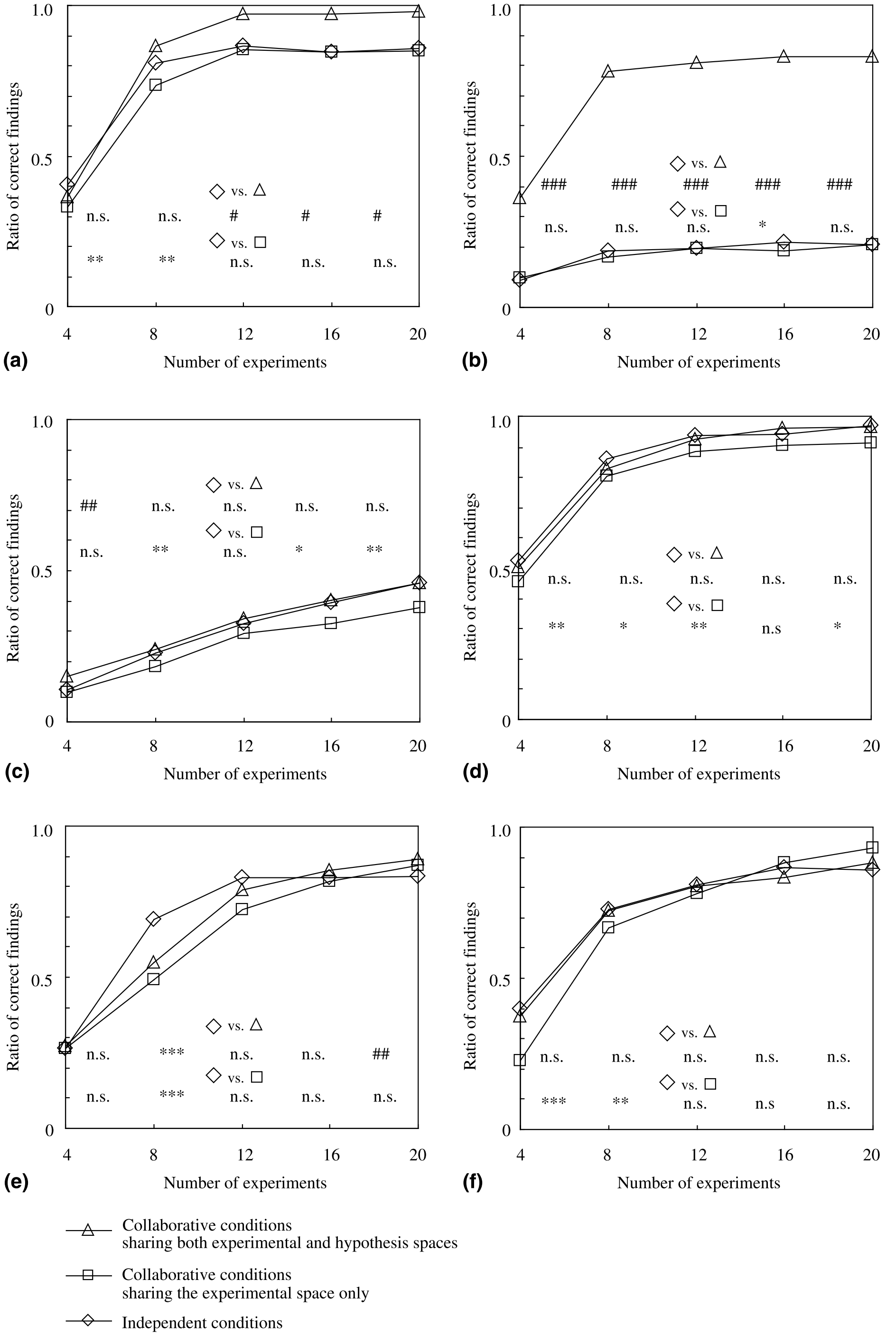


Fig. 4. Results of computer simulations. (The asterisks show the advantage of the independent condition whereas the sharp signs show the advantage of the collaborative condition. The levels of significance were used: ### (or ***) for $p < .01$, ## (or **) for $p < .05$, and # (or *) for $p < .1$. No significance is indicated with n.s.) (a) +Htest vs. +Htest Narrow targets. (b) +Htest vs. +Htest Broad targets. (c) −Htest vs. −Htest Narrow targets. (d) −Htest vs. −Htest Broad targets. (e) +Htest vs. −Htest Narrow targets. (f) +Htest vs. −Htest Broad targets.

In Fig. 4, the performance in the independent condition and that in the collaborative condition are compared. In the independent condition, we regarded the targets as correctly found when at least one of the two systems, each searching independently without interaction, reached the correct solution. Each of these conditions is defined in Fig. 1.

In the collaborative condition, the experiments were alternately conducted (see the example behavior shown in Table 2). Through each simulation, one system generated half of all instances, and the other generated the other half. Each experimental result was shared by both systems, that is, each system knew all generated instances with the Yes or No feedback given to each instance.

The collaborative condition was also subdivided into the following two sub-conditions. In one sub-condition, each system simply alternately conducted the experiments without referring to the hypothesis that the other system had formed. In this sub-condition, the two systems shared only the experimental space. In the other sub-condition, one system tried to form a different hypothesis than that of the other system while referring to the hypothesis of the other system. In the latter sub-condition, the two systems shared the hypothesis space in addition to the experimental space (Klahr & Dunbar, 1988; Simon & Lea, 1974).

In Fig. 4, statistical analysis results are also indicated in the lower portion of the graphs. The upper row shows a comparison between the performance in the independent condition and that in the collaborative condition when the two systems tried to form different hypotheses. On the other hand, the lower row indicates a comparison between the performance in the independent condition and that in the collaborative condition when each system did not refer to the hypothesis of the other system. The asterisks show the advantage of the independent condition whereas the sharp signs show the advantage of the collaborative condition. Three levels of significance were used: ### (or ***) for $p < .01$, ## (or **) for $p < .05$, and # (or *) for $p < .1$. No significance is indicated with n.s.

The comparisons indicate that the performance in the collaborative condition exceeds that in the independent condition, but only when (1) both systems use the +Htest strategy to find broad targets, and (2) both systems try to form different hypotheses while sharing their hypotheses. In the other cases, the effect of collaboration is not significant.

Additionally, in the four settings: (a) +Htest vs. +Htest and Narrow targets, (d) −Htest vs. −Htest and Broad targets, (e) +Htest vs. −Htest and Narrow targets, and (f) +Htest vs. −Htest and Broad targets, the successive increase of “ratios of correct findings” was observed whereas, in the other two conditions: (b) +Htest vs. +Htest and Broad targets and (c) −Htest vs. −Htest and Narrow targets, the performance was relatively low.[1]

As mentioned, the model was regarded as solving a search task, so the model necessarily reaches the solution in the end if it can generate informative instances such as false positives and negative hits. This is the reason for the successive increase in the former four conditions where false positives and negative hits arose many times. On the other hand in the other two conditions the model did not generate such informative instances; that is, the model continuously received positive hits and false negatives. These tendencies are consistent with the predictions by Klayman and Ha’s syntactic analysis.

5. Psychological experiments

To verify the results of the computer simulations described in the previous section, we conducted a psychological experiment.

5.1. Design

A total of 136 subjects participated in the experiment. Based on the experimental setting in the computer simulations, the following two factors were investigated.

[1] These tendencies are more clearly understood when we analyze these data based on three categories of target specificity: extremely narrow targets, medium targets, and extremely broad targets. See Fig. 8 in Appendix A.

5.1.1. Hypothesis-testing strategies

Each of the subjects was assigned to one of the following five experimental conditions: (1) the single +Htest condition where a single participant solved the task using a +Htest, (2) the single −Htest condition, (3) the collaborative +Htest and +Htest condition where two participants, both of whom were required to use a +Htest, collaboratively solved the task, (4) the collaborative −Htest and −Htest condition, and (5) the collaborative +Htest and −Htest condition.

5.1.2. The nature of targets

Each subject solved two problems. In one problem, the subjects had to discover "three evens" as a narrow target. In the other problem, the subjects had to discover "three different numbers" as a broad target. The order of the problems was counterbalanced. Twenty-four trials (experiments) were permitted to find each target. The experimental design is summarized in Table 5.

The subjects were never informed whether they had found a target or not. The information the subjects obtained was only experimental feedback, Yes or No, as a result of their generating an instance. All subjects were requested to generate 24 instances. So some of them were forced to continue to test their hypotheses after they felt they had already found a target. The judgement of their reaching the solution was done by an experimenter. That is, when a hypothesis that a subject forms was identical to a target, the experimenter regarded that he/she found the target at the point; still in this case the subjects were requested to continue their trials until obtaining 24 instances.

In the following discussion, we exclude the subjects who did not follow the experimental instructions requiring the use of each hypothesis-testing strategy. Table 5 shows the number of subjects (or pairs) assigned to each experimental condition; the table also shows, in parentheses, the number of them who correctly followed the hypothesis-testing instructions.

A computerized experimental environment set up on a personal computer was used. When a subject input an instance in an experiment, the system gave a Yes or No feedback to identify whether the instance fit the target. The subject was also required to write, on an experimental sheet, his/her current hypothesis on the target.

In the single condition, a subject individually solved the task. On the other hand, in the collaborative condition, two subjects alternately conducted experiments, and each formed hypotheses while referring to the other's hypotheses (written on the experimental sheet of the partner). The performance in the independent situation, in the psychological experiment here, was calculated from the performance in the single situation, that is, by constructing a virtual pair from the single subjects. The detailed procedure of calculating the independent performance is given in Appendix B.

5.2. Followers and defectors

Before mentioning the main result, we should discuss why relatively many subjects could not follow the experimental instruction. As can be confirmed in Table 5, there are some experimental settings where many defectors emerged.

A unified explanation was not found. However, there are local explanations on some of the distinctive cells in Table 5, where the ratio of defectors is relatively higher than those in other cells.

(1) Single and +Htest and Broad: In this situation, the subjects continuously received positive hits (+Htest and Yes feedback) that did not give them informative evidence. So one explanation of the high ratio in this cell may be that they made

Table 5
Number of participants

	Single		Pair		
	+Htest	−Htest	+Htest vs. +Htest	−Htest vs. −Htest	+Htest vs. −Htest
Narrow	17(15)	18(14)	16(15)	17(11)	15(11)
Broad	17(10)	17(12)	17(12)	17(9)	16(9)

errors by trying to receive informative feedback by using −Htests.

An interesting thing is that this situation also occurs in the pair condition (Pair and +Htest vs. +Htest and Broad). But the ratio of defectors in this pair situation was not so high. As mentioned in the manuscript, in the collaborative situation, −Htest was brought about by the partner's +Htest. This may be the reason for the difference in the ratio of defectors between the single and pair conditions.

(2) Pair and −Htest vs. −Htest and Broad: As many previous studies have suggested, the +Htest is more familiar to human subjects. It is supposed that using −Htests continuously was very peculiar to the participants in this experiment. Additionally, as mentioned in the manuscript, in a pair situation, when at least one of the two subjects made an error, the experimental data were categorized into irregulars. This may be the reason for the many defectors observed in this cell. (However, this interpretation cannot explain why the ratio of defectors in Pair and −Htest vs. −Htest and Narrow was not so high.)

(3) Pair and +Htest vs. −Htest: In this situation, a partner used a different hypothesis-testing strategy. Consequently, each subject would make an error, being influenced by the partner.

5.3. Results

Fig. 5 indicates the experimental results, using the same format as in Fig. 4. Note that the single performance in the +Htest and −Htest combination is absent because this single condition was not set up in the experiment (see Figs. 5(e) and (f)). The independent performance of the +Htest and −Htest combination was calculated from the performance in the single +Htest condition (where the subjects used +Htests only) and the performance in the single −Htest condition, not from the performance in the single +Htest and −Htest condition.

From a statistical analysis, the upper row shows a comparison between the performance in the single condition and that in the collaborative condition, while the lower row shows a comparison between the performance in the independent condition and that in the collaborative condition.

The statistical analysis shows that the performance in the collaborative condition did not exceed that in the independent condition for every combination of hypothesis-testing strategies. The performance in the collaborative condition exceeded that in the single condition only with the combination of +Htest and +Htest to find a broad target. An obvious tendency of the advantage of the collaborative condition over the independent condition was observed even though the statistical analysis did not indicate a significant difference.

As mentioned, the model did not implement the degree of familiarity of each target. So basically, in this study, we cannot compare the absolute performance of human subjects with that of the model. What we can do is to compare the performance in the collaborative condition with that in the independent or single condition in each experimental setting. However, an additional discussion on the difference of performances between the human subjects and the computational model is needed. In (c) −Htest vs. −Htest and Narrow targets, the performance of human subjects is much higher that that of the model. In this situation, the subjects (the model) did not receive much informative evidence because "false negatives" were predominant. The target used in the psychological experiment was "three evens". This target is very familiar to the human subjects. So even though the subjects did not receive informative instances, they could notice the target. On the other hand, in the computer simulations, the model searched the broad hypothesis space one by one to find narrow targets, including "three evens". The model could not reach the solution without negative hits that provide chances for hypothesis revision. This is the reason for the slow increase in performance in the computer simulation.

In the computer simulations, the benefit of collaboration in the +Htest and +Htest combination was confirmed only when each system formed different hypotheses while referring to the other's hypotheses. Therefore, to confirm the effect of two subjects forming different hypotheses, we then conducted the following additional analysis. First, we divided the subjects in each collaborative condition into two groups: subjects who found the correct target earlier (i.e. with fewer instances

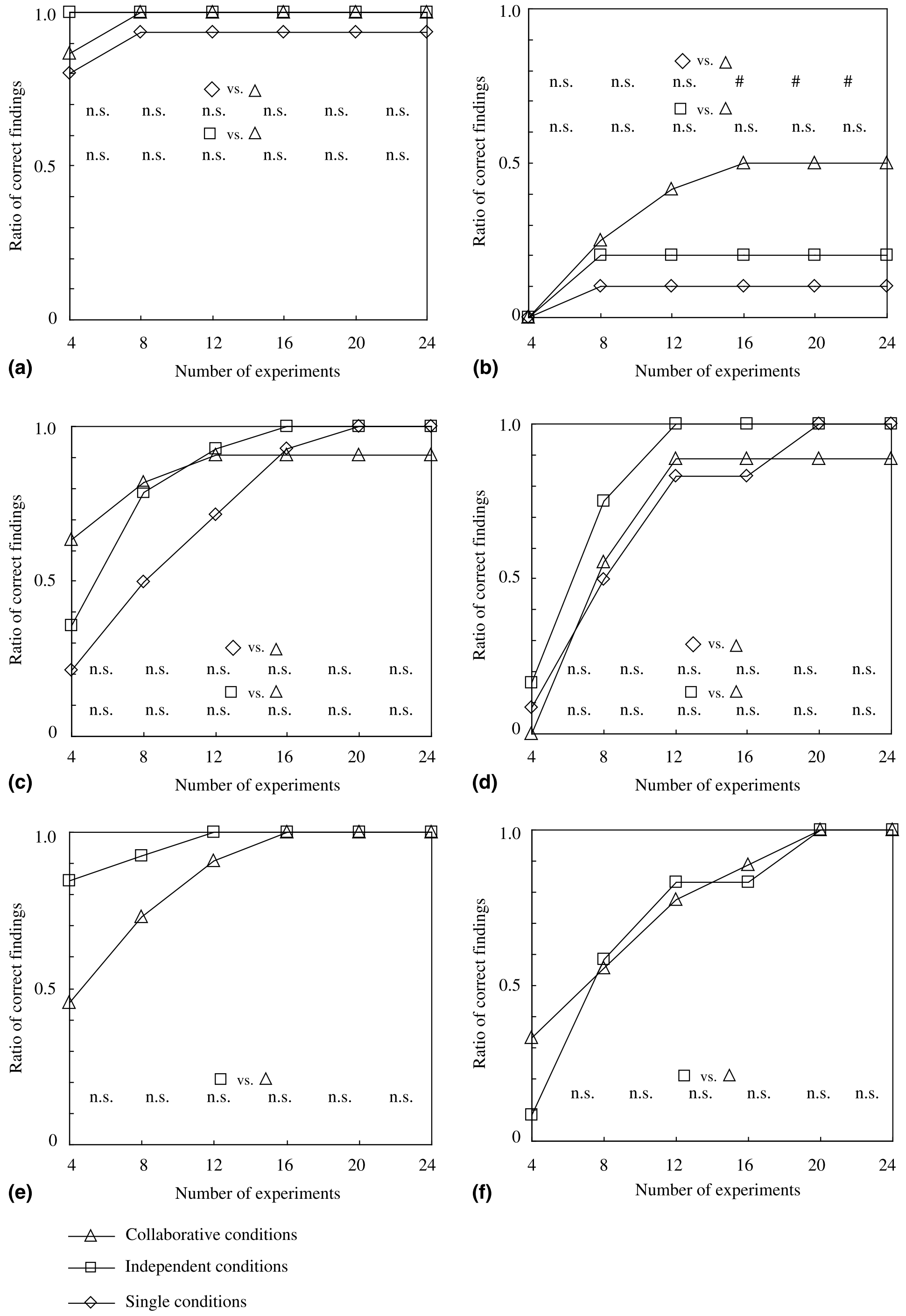


Fig. 5. Results of psychological experiments. (The asterisks show the advantage of the independent condition whereas the sharp signs show the advantage of the collaborative condition. The levels of significance were used: ### (or ***) for $p < .01$, ## (or **) for $p < .05$, and # (or *) for $p < .1$. No significance is indicated with n.s.) (a) +Htest vs. +Htest Narrow targets. (b) +Htest vs. +Htest Broad targets. (c) −Htest vs. −Htest Narrow targets. (d) −Htest vs. −Htest Broad targets. (e) +Htest vs. −Htest Narrow targets. (f) +Htest vs. −Htest Broad targets.

tested) and those who did later. The latter group included those who did not find the correct target. Then, in each group, we averaged the proportions of subjects maintaining different hypotheses throughout the trials (experiments) until reaching the solution. Fig. 6 shows the results. In the +Htest and +Htest combination, the subjects who found the target earlier maintained different hypotheses to a greater extent than the subjects whose performance was lower. It should be noted that the effect of forming different hypotheses appears in the combination of +Htest and +Htest, especially when finding the broad target, whereas this effect does not appear in the combination of −Htest and −Htest. These results are consistent with the findings obtained in the computer simulations.

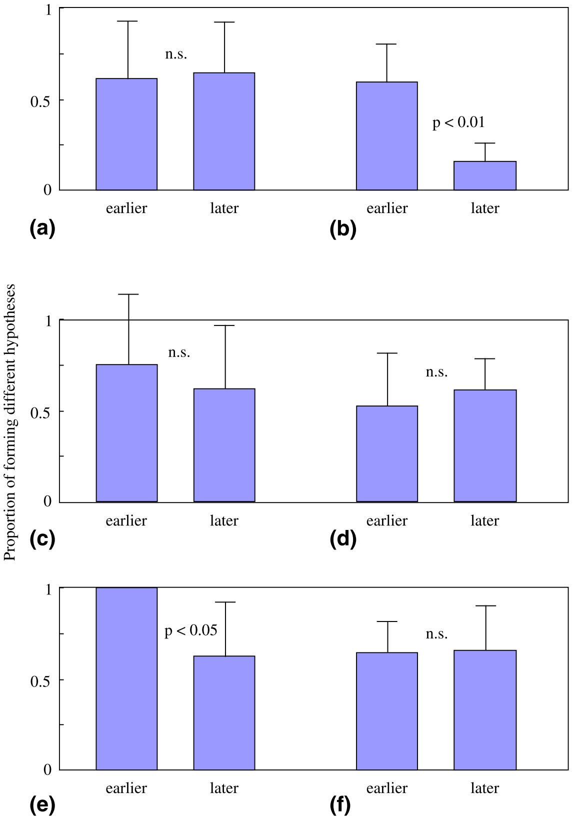


Fig. 6. Averaged proportions of maintaining different hypotheses. (a) +Htest vs. +Htest Narrow targets. (b) +Htest vs. +Htest Broad targets. (c) −Htest vs. −Htest Narrow targets. (d) −Htest vs. −Htest Broad targets. (e) +Htest vs. −Htest Narrow targets. (f) +Htest vs. −Htest Broad targets.

6. Discussions

6.1. Theoretical analysis

Why does the advantage of collaboration emerge only when both systems and both human participants, in finding broad targets, repeatedly conduct a +Htest? We discuss the reason for this based on Klayman and Ha's framework of analysis.

As mentioned before, Klayman and Ha indicated, by their mathematical analysis, that the +Htest was an effective heuristic for finding narrow targets; on the other hand, the +Htest was revealed to be at a disadvantage when finding broad targets (Klayman & Ha, 1987).

Let us consider a collaborative condition in which both of two systems (or two subjects), System A and System B, alternately conduct a +Htest, and each system has a different hypothesis than the other. In this situation, the following accidentally happens: a positive instance for a hypothesis of System A, HA, corresponds to a negative instance for a hypothesis of System B, HB. For example, when hypothesis HA is "the interval is 2" and hypothesis HB is "ascending numbers", the instance is "2, 0, −2" (Fig. 7).

When System A conducts a +Htest using this instance, a −Htest is generated for System B. As a result, the Yes feedback causes conclusive falsification of hypothesis HB because of the combina-

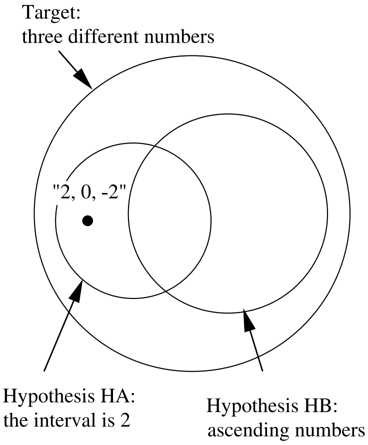


Fig. 7. Situation in which +Htest of System A generates −Htest for System B.

tion of the −Htest for HB and the Yes feedback. This produces the effect of collaboration when two systems, both of which use the +Htest, find broad targets.

An important point is that this function emerges in the interaction between two systems. The advantage of this function is not produced as the effect of the number of the systems. That is, the advantage is not due to the fact that the number of systems in the collaborative condition is twice as many as that in the single condition. As is confirmed in Fig. 7, when each system independently conducts a +Htest, the system's hypothesis is never disconfirmed. The chance of hypothesis falsification can occur only through the collaboration of the two systems.

Twelve pairs in the +Htest and +Htest combination for finding the broad target were analyzed (see Table 5). Actually, 11 among the 12 pairs actually faced the situation discussed above. The examples are shown in Table 6.

The next question is: why does not this kind of effect appear in the combination of −Htest and −Htest when finding narrow targets where the probability of the subjects receiving a No feedback is very high?

If the type of interaction between two systems described above emerges in the combination of −Htest and −Htest, we should see a situation in which the −Htest of System A brings about a +Htest for System B. Generally speaking, however, members (positive instances) of a hypothesis are much fewer than non-members (negative instances). Of course this depends on target rules. As can be confirmed in Table 4, the ratios of the target instances to all instances are relative low. And this is reasonable in a real world context because real-world hypothesis testing most often concerns minority phenomena. This point was also mentioned in (Klayman and Ha, 1987).

Therefore, the possibility of constructing a situation where the −Htest of one system accidentally brings about a +Htest for the other system, resulting in the effect of −Htest and −Htest collaboration, is much lower than the possibility of constructing a situation where the +Htest of one system brings about an −Htest for the other system, resulting in the effect of +Htest and +Htest collaboration. This imbalance between the numbers of positive and negative instances is the reason why only the combination of +Htest and +Htest was observed to produce the effect of collaboration.

We have obtained strong evidence of the effect of collaboration in the +Htest and +Htest combination for finding a broad target because all of the theoretical analyses, computer simulations, and psychological experiments consistently supported this effect.

Table 6
Empirical examples of bringing about a −Htest by the partner's +Htest

Subjects	Instances	Hypothesis HA	Hypothesis HB
S1	8, 5, 2	The interval is the same	The interval is 2
S2	3, 6, 9	The interval is the same	The first is even
	12, 3, 4	The second is less than the sum of the first and the third	Ascending numbers
S3	−8, −10, −12	The interval is the same	Ascending numbers and the interval is the same
S4	−2, 3, 8	The interval is the same	Three evens
S5	1, 3, 6	Positive digits	The interval is the same
	9, 6, 4	Positive digits	Ascending numbers
S6	1, 3, 4	The product is even	The interval is the same
	−39, 99, 5	All digits are less than 100	The sum is less than 50
	111, −10, −2	The sum is less than 100	All digits are less than 100
	−234, 666, 12	The third is less than 100	The sum is less than 100
S7	−4, −3, −2	The interval is the same	The sum is a positive number
S8	6, 8, 2	Three evens	The interval is the same
S9	−2, 0, 6	Three evens	The interval is 2
S10	1, 3, 7	Ascending numbers	The interval is the same
S11	2, 6, 18	Three evens	The interval is the same

The following point should be mentioned. The preceding psychological studies have suggested that human subjects tend to use +Htests. If most subjects use +Htests naturally, as the previous literature has shown, and most of the literature tends to use the more difficult broad targets, then why does the literature not find an advantage for collaboration?

As mentioned in the manuscript, in the experiments shown in Table 1, the subjects' hypothesis tests were not controlled. Even though those subjects conducted mostly +Htests, they were also assumed to conduct a few −Htests. My view was that the results of the −Htests were very informative for solo subjects finding broad targets.

Actually, in the current experiment 7 among 17 single subjects who were instructed to use only +Htests for finding the broad target violated this instruction (see "Single and +Htest and Broad" in Table 5). They actually conducted a few −Htests. However, the performance of those defectors was much higher than that of followers who used only +Htests. That is, the ratio of defectors who found the target was 0.71, whereas the ratio of followers who found the target was only 0.10. This may be the reason why in the preceding studies the advantage of collaboration was not found.

6.2. Collaborative effect without improvement of an individual system's ability

One effect of collaboration that has been widely recognized is the emergence of new cognitive activities or resources in an individual problem solver through interaction among multiple problem solvers. For instance, Okada and Simon made the participants discover scientific laws using a molecular genetics task in a computer micro-world (Okada & Simon, 1997). They found that when the participants collaboratively solved the task, they performed better than when they solved the problem independently.

Through detailed analysis of the participants' discovery processes, it was found that, in the collaborative condition, each of the participants performed deeper explanatory activities brought about by their interaction with the other. This means that the research attributed the cause of the better performance in the collaborative condition to the emergence of a new cognitive activity in an individual participant, which was brought about by the interaction among participants.

Here, we must stress that the effect of collaboration dealt with in the present study is definitely different from the effect characterized in the above studies.

In our simulation, we did not assume that collaboration brings about the emergence of any new mechanism in each individual system. For example, each model's abilities in the collaborative condition, such as the ability to form hypotheses and the capacity of the working memory, are definitely identical to those in the single condition.

In spite of this, the fact that performance in the collaborative condition exceeds that in the independent condition means that even this simple interaction can bring about a kind of emergent phenomenon. More concretely, the findings in the present study imply that even when an individual system cannot improve its abilities through interaction, there is a possibility that the effects of collaboration will emerge.

This kind of effect is important because collaboration does not necessarily provide new cognitive abilities to problem solvers. For example, Laughlin and Hollingshead compared hypotheses generated after conversations among group members with those made before the conversations in the process of problem solving and found only a few emergent hypotheses as a consequence of interaction (Laughlin & Hollingshead, 1995).

6.3. Effects of more than one hypothesis or effects of collaboration

The findings presented in this paper may be understood not as the effects of collaboration but as the effects of simply generating different hypotheses and conducting each experiment by using more than one hypothesis. In fact, some psychological studies have reported that a single subject's performance can be remarkably improved by the experimenter's instructions when the subject is asked to seek two exclusive targets, DAX and MED (Tweney, Doherty, Warner, & Pliske, 1980). The DAX and MED task also caused a shift in the participants' representation of the task that leads to a more effective search strategy, but this issue is not be dealt with in the current paper (Gorman, Stafford, & Gorman, 1987). In this case, the subject is practically led toward generating more than one hypothesis to conduct experiments. Similarly, from the viewpoint of the philosophy of science, the effectiveness of diagnostic tests based on competing hypotheses has been shown (Platt, 1964).

Generally speaking, however, the ability of a single subject or a solo scientist to conduct experiments while generating and maintaining several different hypotheses necessitates huge cognitive and physical costs. In many cases, therefore, it is difficult to do this in practice. Actually, experiments by Freedman et al. using Wason's 2-4-6 task have indicated that, in comparison to discoveries by single subjects, the effects of forming different hypotheses emerge a lot more in group discovery where four subjects, all maintaining different hypotheses, collaboratively work to find a target (Freedman, 1992). They concluded that the reason why more effects emerge in the group condition is because manipulating more than one hypothesis increases the cognitive costs for single-subject problem solving. Note here that in Table 1 the Freedman (1992) study was the only study in which the performance in the collaborative condition exceeded that in the independent condition.

The important finding here is that the collaboration of subjects able to use only restricted cognitive resources, for example, forming only one hypothesis at a time, improves the total performance. This type of collaboration is considered to be one of the important styles of distributed cognition (Hutchins, 1995; Hutchins & Klausen, 1996).

It should be noted that the effect of collaboration appeared only when paired subjects (systems) maintained different hypotheses. When people work together, there is a tendency for their ideas to become more similar as a result of collaborating. Sherif's classic studies demonstrated this effect (Sherif, 1936), and recently Alterman and Garland presented a computational model of this kind of coordination (Alterman & Garland, 2001). This type of effect would work against the positive effect of collaboration, so we should consider a way to compensate for it. For example, groups are more likely to generate creative ideas if the members first work independently (Finke, Ward, & Smith, 1992). This type of combination of independent and collaborative activities may merit further investigation for strengthening the positive effect of collaboration. Additionally, a rule found in Wason's task usually has one dimension, e.g. the number rule. When there are multiple possible dimensions to the rule, as in the New Elusis task, participants sharing information from positive tests may focus together on only one dimension of the task. Exploration on the effects of collaboration in this type of situation may be one of the most important future works.

6.4. Collaboration as interplay producing disconfirmatory information

The importance of negative feedback, such as disconfirmatory instances, to a hypothesis in the process of discovery has been verified in various research fields. For instance, Kulkarni and Simon constructed a production system, called KEKADA, that traced Hans Krebs's discovery process for the elucidation of the chemical pathways for synthesis of urea in the liver (Kulkarni & Simon, 1988). The system forms a hypothesis and predicts the experimental result. An experimental result that significantly differs from the prediction is indexed as a surprising result. They indicated that the surprising result plays an important role in the subsequent discovery process.

Dunbar made a detailed analysis of the research activities in some top-level biological research institutes (Dunbar, 1995, 1997). Again, the importance of surprising results was stressed. He concluded, "the surprising results can be used to generate new hypotheses and research programs". He also regarded unexpected findings observed in the process of investigation as a key concept in discovery and analyzed the reactions of scientists and non-science students to them. He found differences in the method used by the experts and novices for identifying the cause of the unexpected findings (Dunbar, 2001).

In the philosophy of science, the role of anomaly in discovery has been very often discussed (Darden, 1992) and Chinn and Brewer have addressed this issue extensively in their articles (Chinn & Brewer, 1998, 2001). Moreover, in computer science, especially in the fields of concept identification and inductive inference, the importance of counter examples has been generally recognized (Winston, 1992).

All of these concepts, including surprising results, unexpected findings, anomaly, and counter examples, are, in a sense, related to disconfirmatory instances. This fact implies that obtaining high-quality disconfirmatory information is crucial for finding a target.

It is well known that people tend to use positive tests rather than negative tests in various situations (Gorman & Gorman, 1984; Gorman, Gorman, Latta, & Cunningham, 1984; Mynatt, Doherty, & Tweney, 1977; Mynatt, Doherty, & Tweney, 1978). Additionally, a scientist also has a strong positive test strategy (Mahoney & DeMonbruen, 1997). Klayman and Ha pointed out that the positive test strategy is a good all-purpose heuristic in usual contexts.

Let us consider a situation in which a scientist, who has developed a certain incomplete theory that explains only restricted phenomena, tries to generalize his/her theory. This situation often brings about the combination of the use of a +Htest and general (or broad) target finding; therefore, the scientist cannot obtain disconfirmatory instances. This situation prevents the scientist from finding his/her solution. When disconfirmatory instances cannot be generated on a solo basis, those instances should be brought about by another agent.

The results obtained in this research concretely demonstrated an effect of collaboration as interplay producing disconfirmatory instances in this kind of situation. The context set in the present study is highly restricted by using Wason's task. However, the discussion above implies the possibility that the findings here could be effectively applied to more general contexts.

7. Conclusions

In the introduction of this paper, we indicated that the effects of collaboration rarely appear in psychological experiments using orthodox simple discovery tasks. Then we empirically demonstrated a situation in which the effects of collaboration do emerge, and theoretically discussed why such effects appeared. Concretely, we indicated that the effects emerged when both of two subjects (systems) verified their hypotheses by using a +Htest to find broad targets. This finding is more interesting, as a finding on collaborative discovery, when we note that humans have a cognitive bias of tending to use a +Htest more frequently.

From our empirical findings and theoretical discussions, we conclude that (1) generally speaking, simply solving a problem together rarely provides the effects of collaboration, (2) to produce the effects of collaboration, the interaction between collaborative systems must bring about new functions, such as a function for introducing falsification of hypotheses, that are not involved in each individual system, and (3) the possibility of bringing about such abilities depends on the nature of the object that the systems are investigating (such as a target rule in the present study) and the strategies and heuristics that the systems are using (such as a hypothesis-testing strategy), as well as the relationship between these two factors.

Acknowledgements

This research was partially supported by the Ministry of Education, Science, sports and Culture, Grant-in-Aids for Scientific Research (B), 13480094, 2001–2003 and Exploratory Research, 15650177, 2003–2004. I thanks David Klahr (one of three reviewers) for his helpful comments on an earlier version of this paper.

Appendix A

Fig. 8 shows the three-categories analysis where extremely narrow targets are #3, #5, #30, #32, and #33 and extremely broad targets are #19, #27, and #35 shown in Table 4. Other targets are categorized as medium targets.

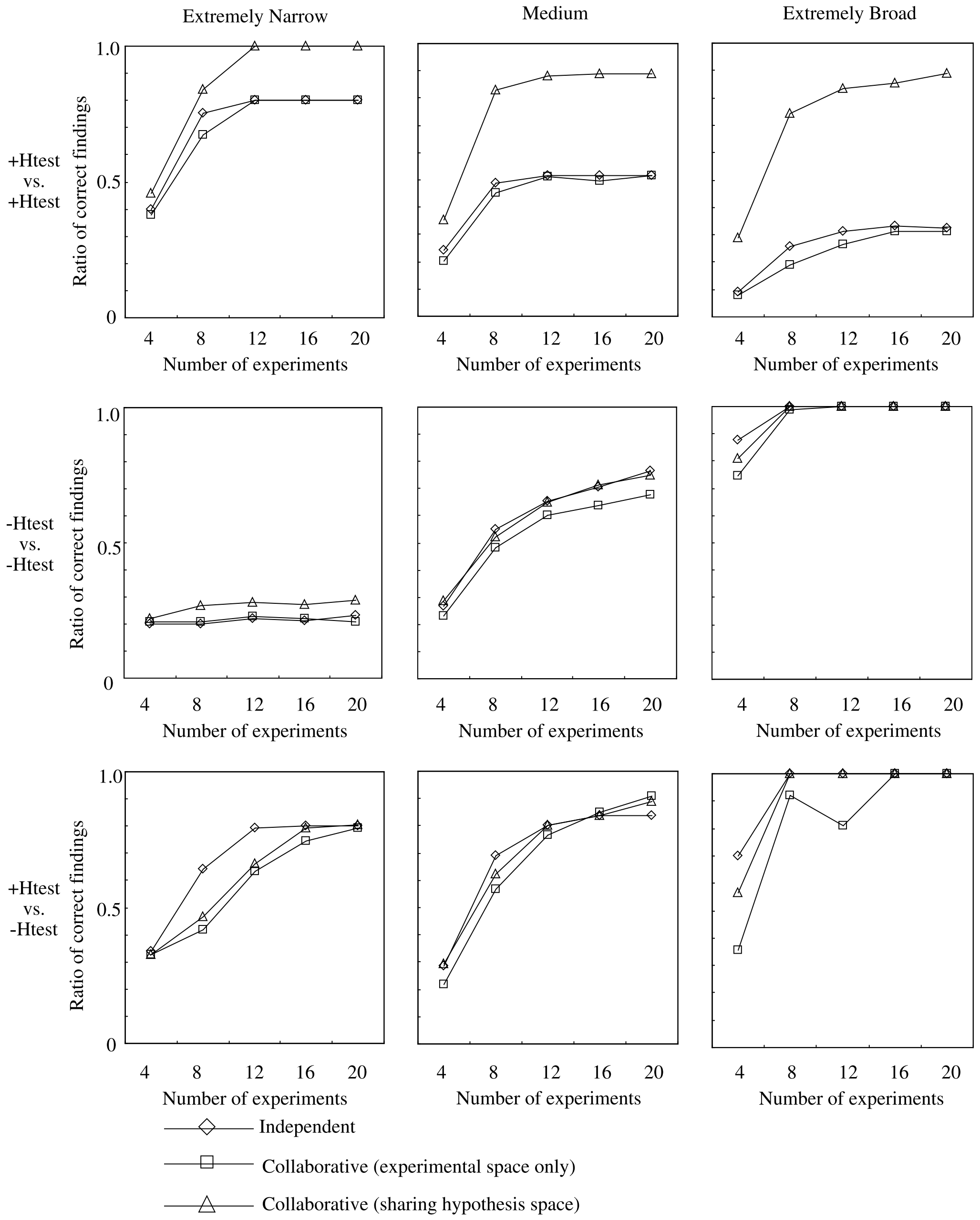


Fig. 8. Three-categories analysis of the results shown in Fig. 4.

Appendix B

The procedure below was used for calculating the performance in the independent condition from that in the single condition.

Let us consider five subjects, each of whom discovers the target when the number of their trials (experiments) reaches 1, 3, 6, 10, and 26, respectively (26 trials means that he/she cannot find the target).

1. First, we create virtual pairs by combining each of the single subjects. In this case, $((5 \times 4)/2)$ pairs are created. The combinations of the number of trials in each pair are the following: $(1, 3)$, $(1, 6)$, $(1, 10)$, $(1, 26)$, $(3, 6)$, $(3, 10)$, $(3, 26)$, $(6, 10)$, $(6, 26)$, and $(10, 26)$.
2. Second, we select the numbers of those who reach the solution earlier. The numbers are the following: 1, 1, 1, 1, 3, 3, 3, 6, 6, and 10.
3. Third, we divide the 10 pairs into five sets in order to match the number of subjects in the independent condition with the number in the single condition. The five sets are the following: 1, 1 | 1, 1 | 3, 3 | 3, 6 | 6, 10.
4. Finally, we use a median of the numbers of trials in each set, rounding up decimals. The medians are the following: 1, 1, 3, 5, and 8. We regard each number as the number of trials when the virtual pairs in the independent condition reach the solution.

References

Alterman, R., & Garland, A. (2001). Convention in joint activity. *Cognitive Science, 25*, 611–657.

Chinn, C., & Brewer, W. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching, 35*, 623–654.

Chinn, C., & Brewer, W. (2001). Models of Data: A Theory of How People Evaluate Data. *Cognition and Instruction, 19*, 323–393.

Darden, L. (1992). Strategies for anomaly resolution. In R. N. Giere (Ed.), *Cognitive models of science. Minnesota studies in the philosophy of science*. Minneapolis: University of Minnesota Press.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight*. Cambridge, MA: MIT Press.

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward (Ed.), *Creative thought: An investigation of conceptual structures and processes*. Washington, DC: American Psychological Association.

Dunbar, K. (2001). What scientific thinking reveals about the nature of cognition. In C. Crowley et al. (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings*. Mahwah, NJ: LEA.

Finke, R., Ward, T., & Smith, S. (1992). *Creative cognition, theory, research, and applications*. Cambridge, MA: MIT Press.

Freedman, E. (1992). Scientific induction: Individual versus group processes and multiple hypotheses. In *Proceedings of the 14th annual meeting of cognitive science society* (pp. 183–188).

Gorman, M. (1992). *Simulating science: Heuristics, mental models, and technoscientific thinking*. Indiana University Press.

Gorman, M. E., & Gorman, M. E. (1984). Comparison of disconfirmatory, confirmatory and control strategies on Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology A, 36*, 629–648.

Gorman, M. E., Gorman, M. E., Latta, R. M., & Cunningham, G. (1984). How disconfirmatory, confirmatory and combined strategies affect group problem solving. *British Journal of Psychology, 75*, 65–97.

Gorman, M., Stafford, A., & Gorman, M. (1987). Disconfiramtion and dual hypotheses on a more difficult version of Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology A, 39*, 1–28.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Hutchins, E., & Klausen, T. (1996). Distributed cognition in an airline cockpit. In D. Middleton & Y. Engestrom (Eds.), *Communication and cognition at work*. Cambridge: Cambridge University Press.

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

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

Klahr, D., & Simon, H. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin, 5*, 524–543.

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

Klayman, J., & Ha, Y.-W. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 596–604.

Kulkarni, D., & Simon, H. A. (1988). The process of scientific discovery: The strategy of experimentation. *Cognitive Science, 13*, 139–176.

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

Laughlin, P. R., & Hollingshead, A. B. (1995). A theory of collective induction. *Organizational Behavior and Human Decision Processes, 61*, 94–107.

Laughlin, P. R., & McGlynn, R. P. (1986). Collective induction: Mutual group and individual influence by exchange of hypotheses and evidence. *Journal of Experimental Social Psychology, 22*, 567–589.

Laughlin, P. R., VanderStoep, S. W., & Hollingshead, A. B. (1991). Collective versus individual induction: Recognition of truth, rejection of error, and collective information processing. *Journal of Personality and Social Psychology, 61*(1), 50–67.

Laughlin, P. R., Bonner, B. L., & Altermatt, T. W. (1998). Collective versus individual induction with single versus multiple hypotheses. *Journal of Personality and Social Psychology, 75*(6), 1481–1489.

Mahoney, M. J., & DeMonbruen, B. G. (1997). Psychology of the scientist: An analysis of problem solving bias. *Cognitive Therapy and Research, 1*, 229–238.

Miwa, K. (1999). Collaborative hypothesis testing process by interacting production systems. *Lecture Notes of Artificial Intelligence, 1721*, 56–67.

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

Miwa, K., Ishii, N., Saito, H., & Nakaike, R. (2002). Changes in learners' exploratory behavior in a simulated psychology laboratory. In *Proceedings of the 24th annual conference of the cognitive science society* (pp. 667–672).

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 24*, 326–329.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology, 30*, 85–96.

Newstead, S.& Evans, J. (Eds.). (1995). *Perspectives on Thinking and Reasoning*. UK: Lawrence Erlbaum Associates Ltd.

Okada, T., & Simon, H. (1997). Collaborative discovery in a scientific domain. *Cognitive Science, 21*, 109–146.

Paulus, P. B. (2000). Groups, teams, and creativity: The creative potential of idea-generating groups. *Applied Psychology: An International Review, 49*(2), 237–262.

Paulus, P. B., & Huei-Chuan, Y. (2000). Idea generation in groups: A basis for creativity in organizations. *Organizational Behavior Human Decision Processes, 82*(1), 76–87.

Platt, J. (1964). Strong inference. *Science, 146*, 347–353.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Sherif, M. (1936). *The psychology of social norms*. New York: Harper.

Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L. Gregg (Ed.), *Knowledge and cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tweney, R. D., Doherty, M. E., Warner, W. J., & Pliske, D. B. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology, 32*, 109–124.

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

Winston, H. P. (1992). *Artificial intelligence* (3rd ed.). Reading, MA: Addison-Wesley.