COGNITIVE SCIENCE Vol 21 (2) 1997, pp. 109–146 ISSN 0364-0213

Collaborative Discovery in a Scientific Domain

TAKESHI OKADA

Nagoya University, Japan

HERBERT A. SIMON

Carnegie Mellon University

This study compares Pairs of subjects with Single subjects in a task of discovering scientific laws with the aid of experiments. Subjects solved a molecular genetics task in a computer micro-world (Dunbar, 1993). Pairs were more successful in discovery than Singles and participated more actively in explanatory activities (i.e., entertaining hypotheses and considering alternative ideas and justifications). Explanatory activities were effective for discovery only when the subjects also conducted crucial experiments. Explanatory activities were facilitated when paired subjects made requests of each other for explanation and focused on them. The study extends from individual to collaborative discovery activities the importance to the discovery process of setting goals to find hypotheses and evidence (Dunbar, 1993) and to construct explanations of phenomena and processes encountered in examples (Chi, Bassok, Lewis, & Glaser, 1989).

Discovery through collaboration is a common and growing practice in science whose processes have not yet been extensively studied. We do not yet know whether discovery processes are different when people work together, what role discussion plays in collaboration, how researchers entertain hypotheses and conduct experiments when working together or how they handle alternative hypotheses and justifications that arise in discussion.

These questions may be approached through historical case studies, field observation, interviews with researchers involved in collaboration, and laboratory experiments. Each approach has its advantages and limitations. This paper follows an experimental approach. We are aware that, although experiments are highly useful for detailed analyses of discovery processes, we need to be cautious in interpreting these data as reflecting actual scientists' collaborative discovery processes. Over the long run, combining various approaches

Direct all correspondence to: Herbert A. Simon, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213.

is most likely to produce a full and accurate picture of collaborative discovery; but one task at a time. In this paper, we first briefly mention historical case data and interview data from our previous studies; then focus on experimental data.

HISTORICAL AND CURRENT COLLABORATION IN SCIENCE

Okada et al. (1995) reviewed three historical cases of collaborative discovery that helped to shape modern science: Jacob and Monod's operon theory in biology; Watson and Crick's double helix in biology; and Simon and Newell's work in artificial intelligence. They identified four features shared by the collaborations. Then, interviewing working cognitive scientists in Japan, they found evidence that each of these features remains essential for successful collaboration: (a) frequent, intense contact between the participants, (b) an egalitarian and exploratory style of discussion, (c) and a shared interest in the research questions, combined with, (d) a diversity in skills and experience (Okada et al., 1995; Schunn, Okada, & Crowley, 1995).

Discussion style is the variable, from among these four, explored in this paper. Although previous surveys suggest the importance of this variable, details of how it affects the process are not available. Our experiment seeks to capture the collaborative processes, paying attention especially to the discussion style.

SOME RELATED RESEARCH ON SCIENTIFIC DISCOVERY

A growing number of studies in cognitive psychology have investigated scientific reasoning and discovery (e.g., Farris & Revlin, 1987; Freedman, 1992; Gorman & Gorman, 1984; Klayman & Ha, 1987; Kulkarni & Simon, 1988; Kuhn & Phelps, 1982; Langley, Simon, Bradshaw, & Zytkow, 1987; Schauble, 1990; Siegler & Libert, 1975; Tschirgi, 1980; Tukey, 1986; Tweney et al., 1980; Wason, 1960). Our work extends to a collaborative setting previous studies that focus upon two processes; hypothesis formation and hypothesis justification (e.g., Dunbar, 1989, 1993; Dunbar & Schunn, 1990; Klahr & Dunbar, 1988; Dunbar & Klahr, 1989; Klahr, Dunbar, & Fay, 1990; Klahr, Fay, & Dunbar, 1993; Qin & Simon, 1990; Teasley, 1995). These studies have tested and extended a dual space model of discovery that was introduced by Simon and Lea (1974) and that provides a basis for integrating the hypothesis formation and testing processes.

According to this model, people search two spaces in discovery: a *hypothesis space* and an *experiment space*. Hypothesis space search builds the structure of a hypothesis and uses prior knowledge or experimental outcomes to assign specific values to its features. Experiment space search tests hypotheses experimentally. Klahr and his colleagues had adults and children conduct experiments to discover the function of a key in controlling a robot vehicle, Big Track. Subjects could form hypotheses and conduct experiments as they wished. Klahr's team found that coordinating the hypothesis space with the experiment space is very important for successful discovery.

Using this framework, Dunbar (1989, 1993) focused on situations where people's previous expectations were disconfirmed by experimental data. Dunbar asked university students to discover how genes are controlled by conducting experiments in a simulated molecular genetics laboratory. Subjects were trained on some elementary concepts of genetics that led them to conclude that a control gene activates the enzyme-producing genes. They were then asked to discover how the enzyme-producing genes are controlled in another simulated genetics model. In this case, the mechanism was inhibition rather than activation. Faced with the unexpected findings, subjects either continued to postulate activation or set a new goal of discovering the cause of the unexpected findings. No subject who adopted the first strategy found the correct answer, but some who adopted the second strategy succeeded.

Although most studies have focused on individual discovery, some have examined collaborative processes (e.g., Freedman, 1992; Gorman, Gorman, Latta, & Cunningham, 1984; Gorman, 1986; Laughlin & Shippy, 1983; Laughlin & Futoran, 1985; Laughlin & McGlynn, 1986; Laughlin, 1988, 1991).

Gorman and colleagues (Gorman et al., 1984; Gorman, 1986) studied confirmation bias in group scientific discovery. For example, they investigated whether groups can falsify hypotheses more effectively than individuals, using a rule-discovery task—the "2-4-6 task." Subjects were instructed to follow either a confirmatory strategy (trying to collect data to confirm their hypotheses), a disconfirmatory strategy, or a combination of the two. Groups performed better than individuals, and subjects in the disconfirmatory condition performed best, followed by those in the combined condition and the confirmatory condition, respectively.

Laughlin et al. observed induction in a group problem solving situation (Laughlin & Shippy, 1983; Laughlin & Futoran, 1985; Laughlin & McGlynn, 1986; Laughlin, 1988, 1991) and showed that both the exchange of hypotheses and of evidence improved performance. These studies suggest that group interaction has an important impact on performance, but they do not report the form and the content of the discussion and how it affected discovery processes.

Teasley (1995) used Klahr, Fay, and Dunbar's (1993) spaceship task, which is similar to the Big Track task, to investigate the role of verbal behavior in children's peer collaborations. Fourth grade students were assigned to one of four conditions: Talk Alones solved the problem alone while talking aloud; No-Talk Alones solved the problem alone without talking aloud; Talk Dyads solved the problem with a partner while talking to each other; No-Talk Dyads solved the problem with a partner without talking to each other. Talk Dyads performed best, followed by Talk Alones and No-Talk Alones, with No-Talk Dyads performing worst. Subjects who produced more interpretive talk that supported reasoning about theories and evidence performed better than subjects who produced less interpretive talk. This study, an important starting point for research on collaborative discovery, leaves many questions unanswered. For example, Teasely (1995) classified statements describing movement of the spaceship as "evidence descriptions." However, we don't know whether or not such descriptions were used for justifying a subject's hypothesis. Teasely also did not focus on alternative hypotheses. Did subjects agree to their partner's interpretations easily, and did they discuss alternative hypotheses and their justifications carefully? Will the findings be replicated with adults?

Collaborative discovery has important aspects that need attention. Klayman and Ha (1989) have shown that successful individual subjects with the 2-4-6 task tended to distinguish explicit alternative hypotheses (diagnostic test strategy). It seems natural to predict that in collaboration, the diagnostic test strategy will also contribute to success. Freedman (1992) studied group versus individual problem solving, including the effects of entertaining multiple hypotheses versus a single hypothesis in the 2-4-6 task. He asked undergraduates to work either individually, or in a four-member group, and to propose either a single hypothesis or a pair of hypotheses. The groups performed better, and in the multiple hypotheses condition, groups used more diagnostic tests than individuals. Freedman concluded that: "individuals may have difficulty forming a mental representation of alternative hypotheses and therefore they are not able to benefit from the presence of multiple hypotheses" (p. 187).

These studies suggest that alternative hypotheses play an important role in collaborative discovery. However, it seems important to study whether, without being forced to entertain alternative hypotheses, collaborators will discuss alternatives and justifications and how such discussion affects discovery.

Some studies have shown that when an experimenter requests subjects to provide explanations, learning improves. When Chi, de Leeuw, Chiu, and LaVancher (1994) asked subjects to explain examples to themselves, the subjects acquired more knowledge. While Chi et al. focused on an individual learning situation, Brown, Palincsar and their colleagues (Brown & Palincsar, 1989; Brown et al., 1991; Palincsar, Brown, & Martin, 1987) have developed a teaching strategy called "reciprocal teaching" which consists of questioning, clarifying, summarizing, and predicting through group discussion, a strategy that includes requests for explanation. The strategy helped students to generate explanations and strongly suggests that Pairs participated in such explanations more often than Singles because Pairs received such requests more often than did Singles.

GENERAL DESCRIPTION OF THIS STUDY

Our focus was on how subjects entertain hypotheses, especially alternative hypotheses, and justify them in a collaborative discovery situation. We sought out the details of discovery processes by using talk-aloud protocols and transcripts of discussions. Such verbal data offer much more detailed and reliable information for understanding discovery processes than do retrospective reports (Ericsson & Simon, 1984). We compared Pairs with Singles, as well as successful Pairs with unsuccessful Pairs, in order to detect important and unique features of successful collaborative discovery.

For several reasons, we used Dunbar's (1993) and Dunbar and Schunn's, (1990) molecular genetics task for this study. First, the mechanism that subjects discover is similar (if much simplified) to the one Jacob and Monod discovered through collaboration. Second, individual discovery processes for this task have been studied previously (Dunbar, 1993; Dunbar & Schunn, 1990), making it easier to compare collaborative with individual processes. Third, as a majority of subjects in previous studies failed to discover the correct mechanism, the task seems to be complicated enough to encourage rich collaboration.

Fourth, in contrast to knowledge-lean tasks such as the 2-4-6 task (Wason, 1960; Klayman & Ha, 1989), subjects who solve this task acquire basic knowledge about the task domain through instruction and practice on a preliminary task, thus making it a more realistic task for research on discovery.

We used as subjects in the Pairs condition friends who were experienced in talking with each other, because the historical case study showed that the collaborators had a close and equal relationship and had spent much time together. Also, previous studies in social psychology showed that subjects in group problem solving situations spent as much time getting to know each other as in solving problems (e.g., Seeger, 1983). Such socialization processes are not the target of this study. Thus, subjects who are already friends provide a more realistic model for scientific collaboration.

Subjects were told at the beginning of the experimental session that they had to report their findings in front of a video camera upon completion of the task so that other subjects could judge the appropriateness of their results. This was intended to make the situation more similar to a real discovery situation in which scientists have to present their findings in public.

Goals

This study aims at describing collaborative discovery processes in detail in terms of hypothesis space search and experiment space search, and specifically, the differences between Singles' and Pairs' discovery processes. We will answer the following questions: (a) Do Pairs perform better than Singles in a scientific discovery task? (b) What are the differences between Pairs' and Singles' discovery processes? (c) What variables are important for success in discovery tasks? We will be especially interested in the effects of collaboration upon the scope and nature of exploration in the hypothesis space and how what is found by search in each space affects the search in the other.

As this study aims at exploring important aspects of collaboration rather than testing a theory, and we wished to analyze the subjects' protocols in detail, we used a relatively small number of subjects in each condition. We report $p < .10$ as statistically significant in this paper so that we can avoid overlooking important features that could otherwise remain hidden.

System Levels

Information processing research has mainly focused on an individual as a cognitive system. This doesn't limit the research to studying internal cognition apart from the environment, for the environment is always present in the stimuli and in the contents of memory (see Vera & Simon, 1993). Nevertheless, such research is concerned with how an *individual* processes information while in interaction with the environment.

Another level of focus is a group of people. Some researchers studying group problem solving in social psychology see a group as a cognitive system (e.g., Laughlin & McGlynn, 1986). Most sociological studies take a community or society as the system to be investigated. More radically, some researchers, like Hutchins (in press), and (Flor & Hutchins,

1991) consider the environment itself as a part of the cognitive system. Similarly, Rogoff (1995) has recently suggested three different levels of analysis: a plane of individual process, a plane of interpersonal process, and a plane of community activity, each level of analysis having its own goals and advantages.

Along with Rogoff, we would argue that the appropriate level depends on what questions we want to answer. In this study we seek to learn whether Singles or Pairs are better able to discover a scientific mechanism, and what contributes to the performance differences. To answer these questions, we compare Singles as cognitive systems with Pairs as cognitive systems.

Data Sources

In the Singles condition of the experiment, the main data are concurrent verbal protocols. It is assumed that these data reflect information that individuals have in working memory during problem solving, although not all information in working memory need be in verbal form, nor is all information in working memory reported in talk-aloud data (Ericsson & Simon, 1984). On the other hand, the main data for the Pairs' condition comes from conversational discourse. These data also do not reflect all of the information that each individual has in working memory during problem solving. It is possible that a member thinks about entirely different things while the partner is talking about his ideas. Even an individual who is speaking about his ideas might not state openly what he is really thinking but might change the content or expression of ideas in order to make himself look smarter or to help the partner understand or accept his ideas. Therefore, in order to compare Singles' and Pairs' verbal protocol data, we must ask in what ways Pairs' discussion data are equivalent to or different from Singles' talk-aloud data.

First, discussion data may not reflect, second by second, the thoughts that occur in each member's working memory. However, we don't need a complete "memory dump," but need mainly to determine whether or not the members attended to the same issues during discovery and reached consensus on their findings.

Second, in order to reach consensus, collaborators must share hypotheses and justifications. If a group member has a hypothesis but does not mention it to the partner, it can not contribute to reaching consensus on the final hypothesis, although unshared ideas might affect the member's generation of new ideas.

Third, one collaborator in a pair might think about topics unrelated to the task while their partner was talking. However, both Pairs and Singles rarely talked about off-task topics.

Fourth, discussion data, as compared to talk-aloud data, might omit important information that led to discovery. However, even in talk-aloud data important information could be missing. Especially when a new idea is emerging in a subject's head, it is hard for researchers to identify its origins because the first glimpse of it often occurs rather suddenly. As we will show in the results section of this paper, however, the hypotheses and justifications in the protocols and discussion transcripts, when accompanied by crucial experiments (also recorded in the data), explained discovery outcomes very well. Therefore, although the analysis might miss some processes, it is likely that the data reflect a large percentage of the processes that influenced individual and collaborative discovery.

Fifth, although the talk-aloud method usually doesn't impair problem solving processes (Ericsson & Simon, 1984), a few recent studies show that in the case of insight problems, verbalization may affect the processes (e.g., Schooler, Ohlsson, & Brooks, 1993). If that is so, the Singles' protocols might not reflect the processes of Singles' problem solving. However, there were no substantial differences of performance between children who talked aloud (Talk Alones) and those who didn't (No-Talk Alones) in previous research in this domain (Teasley, in press). Therefore, the talk-aloud method probably does not markedly change the discovery processes in this kind of task but might slow down the processes somewhat.

Finally, pairs in this study talked more often than Singles about hypotheses, justification, and so on. However, such differences are caused by the nature of collaboration, and we will show that communication between partners provides one key to explaining the advantage of Pairs over Singles in performance on the discovery task.

DESIGN OF EXPERIMENT

We compared a Pairs condition with a Singles condition. In the Pair condition, two subjects collaborated to solve a discovery problem. The Single condition was exactly the same except that the subjects worked alone. Since these two conditions had slightly different procedures from the preceding research, we ran a second singles condition that used the same procedure as Dunbar (1993). The performance of the two Singles conditions was not significantly different (For Dunbar's Singles condition: mean performance score = 1.56 and *SD* = 1.42). Therefore, we don't include Dunbar's Singles condition in the subsequent analyses.

Subjects

Subjects were 27 male undergraduate students at Carnegie Mellon University (CMU) who participated for either course credit or money. Each subject had to: (a) be a male science major undergraduate; (b) bring a friend, who was also a male science major undergraduate, with whom he wanted to participate in collaborative scientific problem solving; and (c) speak English fluently enough to talk aloud or to discuss the problem with the friend. We used science majors mainly because science majors are closer to scientists than non-science majors in educational background and are perhaps more accustomed to scientific thinking. Because it was hard for us to get female subjects to participate and because pilot studies suggested that female Pairs have different discussion styles from male Pairs, we used only male science majors for this study. Obviously, it is important to study female subjects' collaboration in future studies.

After they signed up, subjects were randomly assigned to one of the two conditions. Subjects assigned to the Singles condition were contacted by the experimenter by phone in advance and asked to participate in that condition. We asked each subject if both he and his friend could participate in the experiment separately, to avoid a possibly confounding motivational effect of desire to cooperate with the friend. Eighteen people (nine pairs) were assigned to the Pairs condition and nine people to the Singles condition.

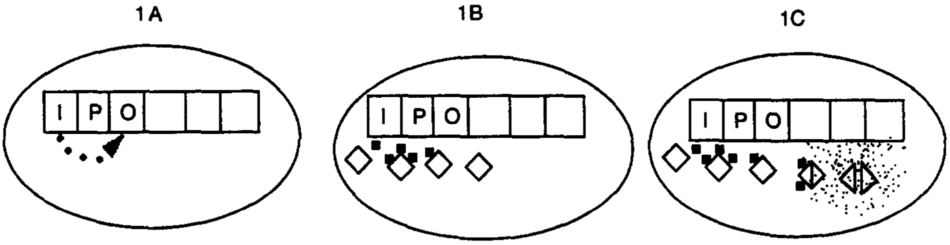
Apparatus and Task

A Macintosh computer running the Simulated Molecular Genetics Laboratory (SMG Laboratory: Dunbar, 1993; Dunbar & Schunn, 1990) was used. In this setting subjects can learn basic concepts and techniques of molecular genetics and conduct simulated experiments to discover a scientific mechanism somewhat similar to the one discovered by Jacob and Monod. A detailed explanation of the mechanism can be found in Dunbar (1993). In this paper, we summarize it briefly.

Genetic Mechanisms and Subjects' Experiments

There are three regulatory genes called I, P, and O. (Although this was not known to the subjects in advance, the P gene is not involved in this regulatory processes.) There are also three genes that produce Beta-gal (an enzyme that breaks down lactose); their production is controlled by the I and O genes. As shown in Figure 1A,[1] in the absence of lactose, the I gene sends chemicals continuously to the O gene, activating it and causing it to block, by physical means, Beta enzyme production. When lactose is present (Figure 1B), the chemicals from the I gene bond with the lactose and do not reach and activate the O gene, thereby permitting the Beta-gal genes to produce Beta enzyme and break down the lactose (Figure 1C). When all of the lactose is broken down (again Figure 1A), the chemicals from the I gene reach the O gene again and reactivate it to inhibit Beta enzyme production. Subjects have to discover that the I gene is a chemical inhibitor and the O gene a physical inhibitor of Beta enzyme production.

The subjects can use two types of experiments for this discovery. One technique is to use mutant genes. As shown in Figure 2B, when I is missing (I mutant), much Beta enzyme was produced. This shows that the I gene inhibited Beta enzyme production when it was



In Figure 1A the E. coli is in an inhibited state: The I gene sends an inhibitor to the O gene, and the inhibitor binds to the O gene, this blocks production of β-gal from the three β-gal producing genes (the three unlabeled genes). In Figure 1B, lactose (diamonds) enters the E. coli. The inhibitor binds to the lactose and not the O gene. In Figure 1C, the β-gal producing genes are no longer inhibited and the beta genes produce β-gal (small dots). The β-gal cleaves the lactose into glucose which can then be utilized as an energy source. When all the lactose has been used up the inhibitor binds to the O gene and the β-gal genes are inhibited from producing β-gal as in Figure 1A.

Figure 1. The cycle of inhibitory regulation of genes in E. coli.

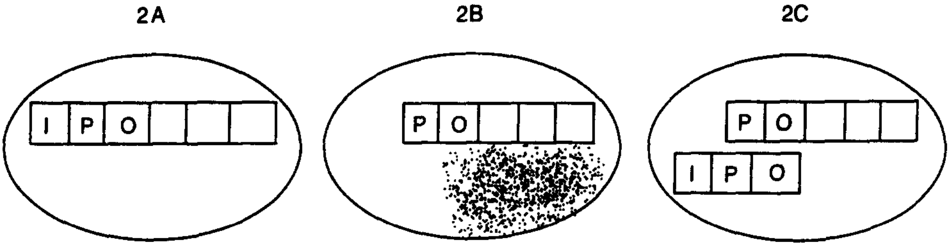


Figure 2A shows a normal haploid E. coli in its resting state, here there is no production of β-gal because the I and O genes inhibit β-gal production. Figure 2B shows a haploid I - mutant, here there is a continous production of β-gal due to the mutant I gene (the I - mutant does not produce an inhibitor). Figure 2C shows a diploid E. coli the female chromosome (on top) is an I - mutant, and the male chromosome (on bottom) is normal, here there is no secretion of β-gal because the I gene on the male chromosome is inhibiting β production on the female chromosome.

Figure 2. Using haploid and diploid E. coli to discover inhibition.

present. Another technique is to use a cell with two chromosomes (See Figure 2C). One is a chromosome with Beta-gal genes (called a female chromosome). The other is a chromosome without them (called a male chromosome). In this case, Beta production was stopped even though the inhibitory I gene is missing from the female chromosome, because the I gene in the male chromosome sends out chemicals to inhibit Beta enzyme production in the female chromosome. Using these two techniques appropriately, the subjects can discover mechanisms similar to those that Jacob and Monod discovered.

The SMG Laboratory (Dunbar, 1993) offers subjects the following choices: six levels of nutrient (0, 100, 200, 300, 400, and 500 units of lactose), three types of mutations (P mutant, I mutant, and O mutant: P- I O, P I- O, and P I O-, respectively) plus normal E. coli (P I O), and two configurations of chromosomes (Haploid E. coli having only female chromosomes or Diploid E. coli having both male and female chromosomes). Combining these dimensions, subjects could conduct 120 types of experiments. Note that Diploid cells could have only one mutant in a chromosome. Hence, there are 96 types of Diploid experiments and 24 types of Haploid experiments.

A video and two audio tape recorders recorded the verbal protocols and the computer display during the discovery sessions. Subjects were asked to discover the mechanism of enzyme production in a cell by conducting experiments in the computer micro-world.

Procedure

The procedure basically followed that of Dunbar (1993). However, some parts were revised in two of the three conditions to adjust the task to the purposes of this study. The procedure consisted of three main phases: (a) Warm-up, (b) Instruction, and (c) Discovery.

Warm-up Phase

To familiarize subjects with giving verbal protocols, they were asked to schedule a list of errands such as going to a movie, picking up a radio, etc. This task was adapted from Hayes-Roth and Hayes-Roth (1979). In the Pairs condition, pairs of subjects were required to discuss and to solve the task together. Subjects in the Singles condition solved it alone while giving verbal protocols.

Instruction Phase

Subjects read the instructions on the computer display in order to acquire basic knowledge and techniques in molecular genetics. They were required to do a practice task on the computer: to find out how A, B, and C genes work to control delta genes to produce delta enzyme in a cell. Because the mechanism of the practice task was activation (i.e., the A gene activates the delta gene to produce delta enzyme), it was expected that subjects would bring the idea of activation into the discovery task as an initial hypothesis. Subjects in the Pairs condition participated in the instruction as pairs, subjects in the Singles condition, as individuals. Subjects could take notes if they wished.

Discovery Phase

In this phase, subjects solved the discovery task. Instructions about the task and the first two experiments were given through the computer display. The subjects' goal was to find out how I, O, and P genes control the beta gene to produce the beta enzyme in a cell. The first two experiments, which all subjects had to conduct as part of the instruction, employed normal haploid cells with 100 lactose and 0 lactose. The results of the two experiments are consistent with the activation hypothesis that most of the subjects generated in the instruction phase. After the first two experiments, the experimenter asked the subjects to report and write down their initial hypotheses and then gave them a summary of the instructions and a copy of the main part of the instructions so that the they could refer to them as needed.

In the Pairs condition, subjects were told: "Now you can do experiments to find out how it really works. Both of you are encouraged to work together in order to reach a consensus on the mechanism. When both of you have reached a consensus on the mechanism, please report the discovery in front of the video camera. Your video-taped report will be reviewed by other subjects who will participate in a similar experiment. They will check whether your ideas are right. You should try to convince them that you are right, using evidence to support your ideas. Remember to work together in order to reach a consensus on the mechanism."

In the Singles condition, subjects were told: "Now you can do experiments to find out how it really works. When you feel that you have discovered the mechanism, please report the discovery in front of the video camera. Your videotaped report will be reviewed by other subjects who will participate in a similar experiment. They will check whether your ideas are right. You should try to convince them that you are right, using evidence to support your ideas. Remember to say everything that you are thinking, everything that is going on in your mind."

Each time subjects conducted an experiment, the computer screen displayed a table with information about their past experiments (type of experiment, amount of lactose input and amount of output). When subjects claimed that they had discovered the mechanism, the experimenter asked them to report their conclusions in front of the video camera and then to write them down.

The Dunbar's Singles condition followed the Dunbar (1993) procedure. The differences between this condition and the Singles condition are that, in the Dunbar's Singles condition, (a) the subjects were not required to report their conclusions in front of the video camera, (b) they were not allowed to take notes during problem solving, (c) they received the warm up task for talk-aloud immediately before the discovery task, (d) they did not receive a copy of the instructions to refer back to, and (e) they were not explicitly taught about the two kinds of effects, chemical and physical. Apart from these differences, the procedure was identical to the other Singles condition.

Hypotheses, Alternative Explanations, Evidence

Three important features of scientific activity will be used as a framework to analyze the verbal protocol data.

1. One goal of science is to build a theory to explain phenomena. Scientists have to think about the explanation as well as the description of a phenomenon. Explanation often takes the form of a hypothesis. Therefore, whether or not people entertain hypotheses is a useful measure in describing the scientific discovery processes.
2. Science progresses through active interaction among members of a scientific community. In order to convince colleagues, a scientist has to consider and discriminate among several plausible alternative explanations. Therefore, the extent to which subjects critique other hypotheses or entertain alternative hypotheses can measure how broadly they search the hypothesis space.
3. In order to convince colleagues, a scientific claim has to be supported by evidence. Scientists have to offer justifications for their arguments or at least to think about what kinds of evidence would support their claims. Therefore, the extent to which subjects consider justifications and the testability of a claim can measure how deeply they search a hypothesis space and how they coordinate hypotheses with data from the experiment space search.

RESULTS AND DISCUSSION

We will report our main results in five parts, that, taken together, provide a coherent picture of the differences in the discovery processes of the Pairs and Singles and the mechanisms producing these differences:

1. After describing our scheme for coding performance (success in discovery), we ask whether or not Pairs performed better than Singles.
2. Having answered this question in the affirmative, we examine eight hypotheses that might explain the superior performance of Pairs and summarize our evaluation of these hypotheses.
3. Having identified explanatory activity as a key variable that discriminates Pairs from Singles, we test its power for predicting performance, and find that explanatory activity is not predictive of discovery unless it is combined with appropriate experimentation.

4. We turn next to examining why Pairs engaged in more explanatory activity than Singles, summing up our answers in Hypothesis 3.1 and 3.2.
5. Finally, examining the relation between hypothesis forming and experimenting, we find that subjects, both Pairs and Singles, divide almost evenly between Theory-Guided Experimenters, who use their hypotheses to plan their experiments, and Empirical Experimenters, who conduct extensive experiments and use the findings to generate their hypotheses. There was no systematic difference in the success of these two strategies in discovery.

Quality of Performance

To determine whether Pairs performed better than Singles, we developed the following coding scheme. Based on their discovery of inhibition and discovery of chemical and physical transmission), subjects' final hypotheses were rated on a 5-point-scale as follows. If a subject did not discover the effect of inhibition at all, we gave 0 points. One point was given to a hypothesis that was correct about only one gene, the I or the O gene.[2] Two points were given to a hypothesis that was correct for both the I and the O genes about inhibition but wrong about chemical and physical transmission. Three points were given to a hypothesis which was correct for both the I and the O genes about inhibition but was correct only for the I or the O gene about chemical and physical transmission. A final hypothesis that described both dimensions correctly (i.e., I was a chemical inhibitor and O was a physical inhibitor) received 4 points. This coding scheme was different from Dunbar's (1993). In Dunbar's study, many of the subjects failed to discover inhibition, and he did not pay attention to the chemical and physical transmission which was discovered afterward. In our study, all Pairs and some Singles discovered inhibition. Therefore, we also focused on the chemical and physical transmission.

Another coder was taught this scheme and coded all of the performance data, except two cases, independently. The second coder used the two cases to practice coding and received feedback on accuracy before starting the coding process. The percentage of consistency between the two coders was 75%. However, since the scores were based on a 5 point scale, we also calculated the correlation between the two coders' scores, finding it to be very high and statistically significant ($r(1, 14) = .93, p < .001$).

Did Pairs Perform Better Than Singles?

Yes. We conducted a one-sided t-test to compare these two conditions. As Table 1 shows, Pairs outperformed Singles ($t(16) = 2.69, p < .01$), with mean scores of 2.89 and 1.67, respectively. A U-test for rank order was also conducted with a closely similar result ($U = 13.5, p < .05$).

In order to check whether there were differences between Pairs' and Singles' intelligence and initial knowledge, we compared Pairs and Singles in terms of SAT scores and initial hypotheses. We asked the subjects to report their SAT scores but obtained them from only 5 Pairs and 6 Singles. These scores do not serve as a complete measure of the subjects' intelligence, but we didn't find any differences between Pairs and Singles in terms of reported scores (See Table 1). All Singles and all except one Pair reported an activation

TABLE 1
Differences Between Pairs and Singles

Measures	Pairs *Means and (SDs)*	Singles *Means and (SDs)*	*p* of *t* Tests
Discovery score (full time)	2.89 (0.93)	1.67 (1.00)	<.05
Discovery score (at 23.02 minutes: Singles' average time)	2.33 (1.23	1.67 (1.00)	=.22
Hypothetical Pairs' (see Singles' column) discovery score	2.89 (0.93)	2.11 (0.67)	=.03
Reported Math SAT scores			
(Pairs -> Average)	706.25 (30.38)	663.33 (97.97)	=.43
(Pairs -> Higher score)	738.00 (23.87)	— —	=.13
Reported SAT scores (Math + Verbal)			
(Pairs -> Average)	1253.00 (23.87)	1246.67 (127.85)	=.92
(Pairs -> Higher score	1310.00 (46.90)	— —	=.32
Solution time (min.)	29.33 (14.85)	23.02 (10.98)	=.32

hypothesis (a wrong hypothesis in a wrong frame) as their initial hypothesis. Therefore, the difference in performance was unlikely to be caused by differences in initial knowledge or intelligence.

Reasons for Superiority of Pairs Over Singles

Why did Pairs perform better than Singles? In this section, we list and test plausible hypotheses.

Hypothesis (1-1): Pairs simply had twice as great a chance as Singles of getting the right hypothesis, even without active interaction.

To test this possibility, we paired subjects in the Singles' condition in all combinations (i.e., 9 × 9 cases) as though we had two sets of nine subjects each with identical scores. In each case we calculated, from the original coding of the hypotheses that the members of the pair had found, the combined score of the pair (the totality of distinct hypotheses found by the two), exactly as we had done for real pairs. Then we calculated the mean and *SD* of the Hypothetical Pairs (still using 8 degrees of freedom, as there were only 9 independent observations) and compared them with the real Pairs' mean and *SD* (See Table 1). The mean score of the real Pairs (2.89) was significantly better than that of the Hypothetical Pairs (2.11) ($t(16) = 1.95, p = .035$). This result indicates that the superior performance of Pairs depends on the members' interactive processes and not just the performance of the abler member.

Hypothesis (1-2): Pairs spent more time than Singles. Only time matters.

If Pairs had been more motivated and spent more time on the task than Singles, this might account for the difference. However, the difference in solution time between the two conditions, shown in Table 1, is not statistically significant ($t(16) = 1.03, p = .32$).

Hypothesis (1-3): Pairs searched the experiment space more effectively than Singles.

To test this hypothesis, we investigated subjects' experiment space search processes using four sorts of measures adapted from Schunn and Dunbar (submitted).

1. Number of Experiments

This measure tests if the sheer number of experiments was a better predictor of success than the content of experiments.

2. Breadth of the Experiment Space Search

These measures focus on how broadly the subjects searched the experiment space, and include (a) Dimensions Searched, (b) Percentage of Genes Searched, (c) Amount of Lactose Searched, and (d) Number of Experiments with Zero Lactose. The Dimensions Searched score was calculated by the following formula: The different kinds of haploid mutations examined (out of four possible patterns: I-, O-, P-, and Normal), plus the number of different mutations in the *first* chromosome in a *diploid* cell examined (out of four possible patterns: I-, O-, P-, and Normal), plus the number of different mutations in the *second* chromosome in a *diploid* cell examined (out of four possible patterns: I-, O-, P-, and Normal), plus the number of different amounts of lactose used (out of 6 possible patterns). The maximum score is 18. This is the most comprehensive measure of number of dimensions searched. The other three measures served as sub-categories to capture breadth of search.

3. Informativeness of Experiments

This category includes (a) the percentage of possible types of crucial experiments that were conducted (out of five types), (b) total number of crucial experiments, and (c) total number of non-crucial experiments. The crucial experiments, those that were necessary to discover the correct mechanism, involve three haploid mutants (I-, P-, and O-) and two diploid mutants. One diploid has the first chromosome with I- mutant and the second chromosome with genes other than I- mutant (*diploid I- crucial experiment*); the other diploid has the first chromosome with O- mutant and the second chromosome with genes other than O-mutant (*diploid O- crucial experiment*). The haploid experiments were necessary and sufficient to discover the inhibitory function of the I gene and the O gene. The diploid experiments were necessary to discover that the I gene has a chemical effect, and the O gene, a physical effect.

4. Systematic Search in the Experiment Space

This category employs the mean feature difference score, which measures how systematically subjects conducted experiments, Varying one variable at a time (called *Vary One Thing at A Time* [VOTAT] by Tschirgi, 1980) is sometimes regarded as fundamental to the experimental method (e.g., Tschirgi, 1980; Kuhn & Phelps, 1982). If several variables are changed at once, it is hard to know which was responsible for the result. The score is the mean number of features subjects changed between two adjacent experiments. Experiments are regarded as less informative as more features are changed.

TABLE 2
Differences Between Pairs and Singles: Experiment Measures

	Measures	Pairs' Mean and (SD)	Singles' Mean and (SD)	p of t-tests
Activity	Number of experiments	13.89 (7.54)	13.89 (6.92)	N.S.
Breadth of E-space search	Dimension search score	11.78 (2.33)	11.44 (2.55)	N.S. (=.78)
	% of genes searched	47.22 (21.08)	41.67 (15.00)	N.S. (=.53)
	% of amounts of lactose searched	46.30 (24.70)	46.30 (18.20)	N.S.
	Number of experiments with zero lactose	3.11 (2.26)	4.56 (4.42)	N.S. (=.40)
Crucial experiments	% of types of crucial experiments	88.89 (10.54)	86.67 (14.14)	N.S. (=.71)
	Number of crucial experiments	10.33 (4.82)	9.67 (4.82)	N.S. (=.77)
	Number of noncrucial experiments	3.56 (2.96)	4.22 (2.95)	N.S (=.64)
	Haploid crucial experiment with I-	1.00 (0.00)	1.00 (0.00)	N.S.
	Haploid crucial experiment, P-	1.00 (0.00)	1.00 (0.00)	N.S.
	Haploid crucial experiment, O-	1.00 (0.00)	1.00 (0.00)	N.S.
	Diploid crucial experiment, I-/N	1.00 (0.00)	0.78 (0.44)	N.S. (=.15)
	Diploid crucial experiment, O-/N	0.44 (.53)	0.56 (.53)	N.S. (=.66)
Systematic search (VOTAT)	Mean feature difference score	1.83 (.14)	1.70 (.29)	N.S. (=.26)

Table 2 shows the means and SDs for each measure. There were no significant differences between Pairs and Singles in experiment space search. We also divided the subjects' solution time into two periods: first half and second half, and compared Pairs and Singles in each period. Again, no statistically significant differences were found. Overall, experiment space search does not show any difference between Pairs and Singles.

Hypothesis (1-4): Pairs entertained hypotheses more often than Singles.

To test the possibility that Pairs performed better than Singles because they talked more often about hypotheses, we created two measures. The first is the *percent of units* in which subjects entertained at least one hypothesis. We define a unit as the period between two adjacent experiments and also the discussion period following completion of the final experiment. We chose this interval for the following reasons: (a) There was no feedback from experimental outcomes between the two adjacent experiments; (b) Analyses using this unit showed significant differences between Pairs and Singles and also explained performance; (c) This unit offers information of the relation of verbal protocols to experimentation; (d) Practically speaking, this is a manageable level to analyze the data in this study;

and (e) Teasely (1995) showed that the results of analyses at this level showed the same pattern of behavior as analyses of protocols sentence by sentence.

The definition of a hypothesis is given in Appendix 1. This Appendix also provides definitions of many other categories of protocol behavior. In this paper, our special focus is on hypotheses, alternative hypotheses, justifications, and statements that enhance explanation. Although we also coded some other categories, we do not describe those data here either because they are irrelevant for the purpose of this study, or because they did not show any difference between Pairs and Singles.

The second measure is the *absolute number of hypotheses* that the subjects entertained throughout the entire discovery process. Hypotheses are described as combinations of variables (such as genes) and functions (such as inhibitors and activators). The hypotheses that subjects mentioned were counted. Appendix 2 shows an example of protocols and coding of hypotheses and other measures.

In order to check the reliability of the coding categories, one coder labeled all the protocols first. Then, another coder labeled one Single's and one Pair's protocols, which were picked randomly. The second coder was taught the definitions and examples of each categories with Appendix 1 and then labeled one Pair's and one Single's protocol for all categories in Appendix 1 as practice. When the second coder miscoded, she got failure feedback and was told why it was wrong. Then the second coder coded the target protocols independently. Percentage of consistency between two coders varied from 72% to 100%, depending on categories in Appendix 1. The average score for all categories was 94%.

Pairs entertained hypotheses more often than Singles. Table 3 shows the percentage of units in which subjects entertained hypotheses. Although the difference (74% vs. 56%) was not significant, it approached the 10% level ($t(16) = 1.74, p = .10$). Table 3 also shows the number of hypotheses that subjects entertained while working on the task. Pairs entertained about twice as many hypotheses as Singles (29.56 vs. 14.00; $t(16) = 3.25, p < .01$). The difference was more salient when we checked the hypotheses in the first half period ($t(16) = 4.81, p < .001$). Pairs entertained about three times as many hypotheses as Singles in the first half of their sessions (14.56 vs. 4.67).

Hypothesis (1-5): Pairs entertained alternative hypotheses more often than Singles.

Recent research on discovery found that forcing people to report alternative hypotheses each time they conduct an experiment can facilitate performance on discovery tasks (Freedman, 1992). However, in Freedman's setting subjects were forced to report two hypotheses every time they conducted an experiment. Subjects do not necessarily entertain alternative hypotheses for all the experiments they encounter when they are learning from examples Okada (1992). In the discovery task, subjects might sometimes consider alternative hypotheses occasionally, but not always. Therefore, we need to check how often the subjects in each condition entertained alternative hypotheses.

Two measures were created. One is the *percent of units* in which subjects entertained *alternative hypotheses*. As defined in Appendix 1, when the subject (s) mentioned two different hypotheses about a variable in one unit, the unit was coded as having an alternative hypothesis. The other measure is the *absolute number of different types of hypotheses* that

TABLE 3
Differences Between Pairs and Singles: Protocol Measures

Measures	Pairs *Means* and *(SDs)*	Singles *Means* and *(SDs)*	*p* of *t* tests
Number of words	2216.22 (1135.41)	1090.67 (489.90)	<.05
First half	1073.56 (573.72)	475.22 (226.82)	<.05
Second half	1142.67 (569.33)	615.44 (271.51)	<.05
Number of hypotheses	29.56 (13.45)	14.00 (5.10)	<.01
First half	14.56 (5.59)	4.67 (2.60)	<.01
Second half	15.00 (9.27)	9.33 (3.43)	=.10
Hypothetical Pairs' (see Singles' column) number of hypotheses	29.56 (13.45)	28.00 (6.45)	=.61
Number of different types of hypotheses	10.44 (3.24)	7.78 (2.28)	=.06
First half	7.11 (2.76)	2.89 (1.36)	<.01
Second half	6.11 (3.55)	6.22 (1.72)	=.93
Hypothetical Pairs' (see Singles' column) number of types of hypotheses	10.44 (3.24)	13.11 (2.69)	<.05
% of units with summarizing data	48 (21)	47 (13)	=.89
% of units with hypotheses	74 (21)	56 (24)	=.10
% of units with prediction	31 (19)	14 (13)	<.05
% of units with extension	37 (15)	22 (14)	<.05
% of units with critque	33 (21)	2 (4)	<.01
% of units with suspension	10 (14)	2 (3)	=.08
% of units with alternative hypotheses	25 (26)	6 (10)	=.05
% of units with combined-justification	58 (29)	39 (24)	=.15
% of units with justification through experimental results	41 (19)	24 (11)	<.05
% of units with plan for new experiments to test hypotheses	37 (24)	17 (16)	=.05
% of units with testability of hypotheses	9 (8)	2 (6)	<.05
% of units with justification using several experimental results	35 (19)	19 (13)	=.05
% of units with argument about justification	24 (16)	9 (10)	<.05
Hypothetical Pairs' (see Singles' column) % of units with hypotheses	74 (21)	70 (21)	=.58
Hypothetical Pairs' % of units with alternative hypotheses	25 (26)	11 (11)	<.05
Hypothetical Pairs % of units with combined-justification	58 (29)	52 (22)	=.49
Hypothetical Pairs' % of units with justification through experimental results	41 (19)	31 (9)	<.05
Hypothetical Pairs' % of units with plan for new experiments	37 (24)	26 (14)	=.08
Hypothetical Pairs' % of units with testability of hypotheses	9 (8)	4 (7)	=.06
Hypothetical Pairs' % of units with justification using several results	35 (19)	26 (14)	=.11
Hypothetical Pairs' % of units with argument about justification	24 (16)	15 (10)	<.05

subjects entertained throughout the entire discovery process. This measure offers a picture of how large an area of the hypothesis space subjects covered.

Hypothesis (1-6): Pairs talked about justification more often than Singles.

Several measures of justification were created. One is the *percentage of units* in which subjects talked about *justification from experimental results.* This measure captures whether subjects showed data to support their hypotheses. Another measure is the *percentage of units* in which subjects talked about *justification using several experimental results.* Klahr and Dunbar (1988) showed that there were differences on this measure between good performers and poor performers in their experiment.

If subjects don't have enough experimental data to support their current hypothesis, one response is to plan a new experiment. Therefore, a third measure is the *percentage of units* in which subjects *planned a new experiment* to test a hypothesis.

If subjects don't have enough experimental data to support their current hypotheses, they may consider if and how the hypotheses can be tested. Therefore, a fourth measure is the *percentage of units* in which subjects talked about *the testability of hypotheses.* A fifth is the *percentage of units* in which subjects *argued against a justification* (partner's or own).

As Table 3 shows, all of the measures that we described above indicate that Pairs considered justification of their hypotheses more often than Singles (justification with data: 41 vs. 24; $t(16) = 1.89, p < .05$; justification with several experimental results: 35 vs. 19; $t(16) = 2.08, p = .05$; experiment to test a hypothesis: 37 vs. 17; $t(16) = 2.07, p = .05$; Testability: 9 vs. 2; $t(16) = 2.21, p < .05$).

Pairs may have entertained justifications more often than Singles simply because Pairs entertained hypotheses more than Singles. We also compared the scores between the two groups by the percentage of units with various measures of justifications divided by percentage of units with hypotheses. The results showed the same pattern of findings as shown above. In general, Pairs entertained justifications more often than Singles (Justification with results: Pairs 58% vs. Singles 44%, $t(16) = 1.39, p = .18$; Justification with several results: Pairs 48% vs. Singles 36%, $t(16) = 1.17, p = .26$; Plan for new experiments: 51% vs. 30%, $t(16) = 1.70, p = .11$; Testability: Pairs 11% vs. Singles 3%, $t(16) = 2.08, p = .05$; Argument about justification: Pairs 33% vs. Singles 13%. $t(16) = 2.46, p < .05$).

We also combined three main measures (i.e., justification with results; new experiment to test a hypothesis; testability of hypotheses) to form a summary measure of justification. When subjects already have data to justify their hypotheses, they just need to mention the data. When they don't have the data yet, but it seems easy to get them, they need to plan new experiments to gather the data. When it seems very hard or impossible to get the data, they might talk about whether they can test the hypotheses. Therefore, these three measures are functionally equivalent. If subjects had one of those in a unit, the unit was marked as Combined-Justification. Table 3 also shows the results. Although the difference was not statistically significant, Pairs considered justification more often than Singles (58 vs. 39; $t(16) = 1.60, p = .13$).

Overall, data suggest that an important reason why Pairs performed better than Singles is because Pairs participated in explanatory activities such as entertaining hypotheses (hypothesis 1.4), talking about alternative ideas (hypothesis 1.5), and considering justification (hypothesis 1.6) more often than Singles.

Variables That Affect Performance

So far, we identified the differences between Pairs and Singles that might cause the differences in performance. In order to support this inference, we should inquire how measures such as entertaining hypotheses and thinking about justification predict performance. Although the number of subjects in each condition is not large enough to conduct multiple regression analysis with many measures, exploratory regression analyses can give a sense of which variables were important for discovery. Therefore, we conducted simple regression analyses between performance and each measure in Table 2 and Table 3 in each condition separately.

We had predicted from the previous analyses that entertaining hypotheses, considering alternatives, and thinking about justification would explain performance well. Contrary to this prediction, there was no significant correlation between performance and these verbal protocol measures (hypotheses, alternatives, and justification). Instead, there were strong predictors of Pairs' performance in the experiment space search measures, which did not predict Singles' performance at all. The strongest predictor for Pairs' performance was percentage of crucial experiments. As we said in the previous section, five crucial experiments are necessary and sufficient to make the correct discovery. This measure accounted for 78% of the variance in Pairs' performance ($F(1, 7) = 24.11, p < .01$), but only 3% of the variance in Singles' performance ($F(1, 7) = 0.23, p = .65$). Other experimental measures that were strong predictors for performance of Pairs include percentage of amounts of lactose searched (Pairs: $F(1, 7) = 7.02, p < .05, r^2 = .50$; Singles: $F(1, 7) = .72, p = .42, r^2 = .09$) and number of experiments with zero lactose (Pairs: $F(1, 7) = 7.61, p < .05, r^2 = .52$; Singles: $F(1, 7) = .11, p = .75, r^2 = .02$).

The results from verbal protocol measures suggest that entertaining hypotheses and considering justification are important measures to distinguish Pairs and Singles. However, the results from the regression analyses suggest that the method of conducting experiments is far more important than entertaining and justifying hypotheses for discovery. To reconcile these apparently conflicting findings, we propose the following interpretation:

> *Hypothesis (1-7): Due to their active participation in explanatory activities such as entertaining hypotheses and considering justification, Pairs could use information from experiment space search, especially information from crucial experiments, effectively in order to make discoveries. On the other hand, Singles could not do so because they did not actively participate in explanatory activities.*[3]

To check this possibility, we divided subjects according to their scores for percentage of crucial experiments, and their mean scores for explanatory activities (i.e., the combined score of entertaining hypothesis and thinking about justification). For the percentage of crucial experiments score, we divided the subjects depending on whether they conducted all 5 types (i.e., 100%) of crucial experiments or not. For the explanatory activity score, we

TABLE 4
Means (SDs) of Performance Scores According to Occurrence of Crucial Experiments and Explanatory Activities

Crucial	Explanatory Activities		
Experiments	High	Low	Totals
High	3.75 (0.50) (n = 4: Pairs 3, Singles 1)	2.00 (1.41) (n = 4: Pairs 1, Singles 3)	2.88
Low	1.75 (0.50) (n = 4: Pairs 2, Singles 2)	1.83 (0.75) (n = 6: Pairs 3, Singles 3)	1.80
Totals	2.75	1.90	

counted as "high" subjects whose scores of both entertaining hypotheses and considering justification are higher than the average of those who had high explanatory activities.

Table 4 shows means and SDs for performance in each cell. Due to the small number of subjects, Pairs and Singles were combined for this analysis. This table shows that the subjects who both conducted all of the crucial experiments *and* actively participated in explanatory activities outperformed the subjects who did only crucial experiments, only explanatory activities, or neither. An ANOVA shows significant main effects (crucial experiments: $F(1, 14) = 6.94, p < .05$; explanatory activities: $F(1,14) = 4.11, p = .06$); and interaction: ($F(1,14) = 4.97, p < .05$). Subjects needed to participate actively in both crucial experiments and explanatory activities in order to discover the right mechanism. Neither one, by itself, was enough. Three of the nine Pairs achieved this combination, but only one of the nine Singles.

These data show clearly that entertaining hypotheses and thinking about their justification play quite important roles in discovery, especially when the experiments are informative. The data do not imply that explanatory activities, *caused* more crucial experiments to be generated, but rather that they assisted in making good use of the findings. However, before reaching a final conclusion about the role of explanatory activities, we should consider the following hypothesis:

> *Hypothesis (1-8): Two Singles could produce as much explanatory activity as one Pair.*

This hypothesis raises the issue of whether merely engaging in explanatory activities is sufficient or whether collaborative explanatory activities are necessary. To test this hypothesis, we carried out analyses of discussion processes similar to those on performance testing hypotheses (1-1). As Table 3 shows, on comparing the real Pairs and Hypothetical Pairs, we found no difference between those two groups in terms of number of hypotheses generated. As to number of types of hypotheses, the Hypothetical Pairs were even better than the real Pairs. We also checked other discussion measures. We could not compare these two conditions directly in terms of measures that used units for analyses because, in case of Hypothetical Pairs, the subjects had different numbers of units. Therefore, we adopted the better score of two Singles as a Hypothetical Pair's score. Although it is not a perfect measure, it offers useful information about the two-man-power hypothesis. As

Table 3 shows, in general, the real Pairs entertained alternative hypotheses and participated in justifications more often than the Hypothetical Pairs. As the performance of the real Pairs was better than that of the Hypothetical Pairs, these data suggest that just entertaining many hypotheses or many different hypotheses is not sufficient for discovery. Instead, it appears that *interactive* or *collaborative* explanatory activities, especially on alternative hypotheses and justifications, are important. This is still a conjecture: we need further research to test the hypothesis more thoroughly.

Reasons for Differences in Explanatory Activity

Why Did Pairs Entertain Hypotheses and Justifications More Often Than Singles?

We have shown that Pairs actively participated in explanatory activities more often than Singles. We will now try to account for this difference. As Figure 3 shows, scientific explanations have various levels. From the shallower level to the deeper, they move from a mere description of results, limited to the specific case; to a summary, which includes a general-

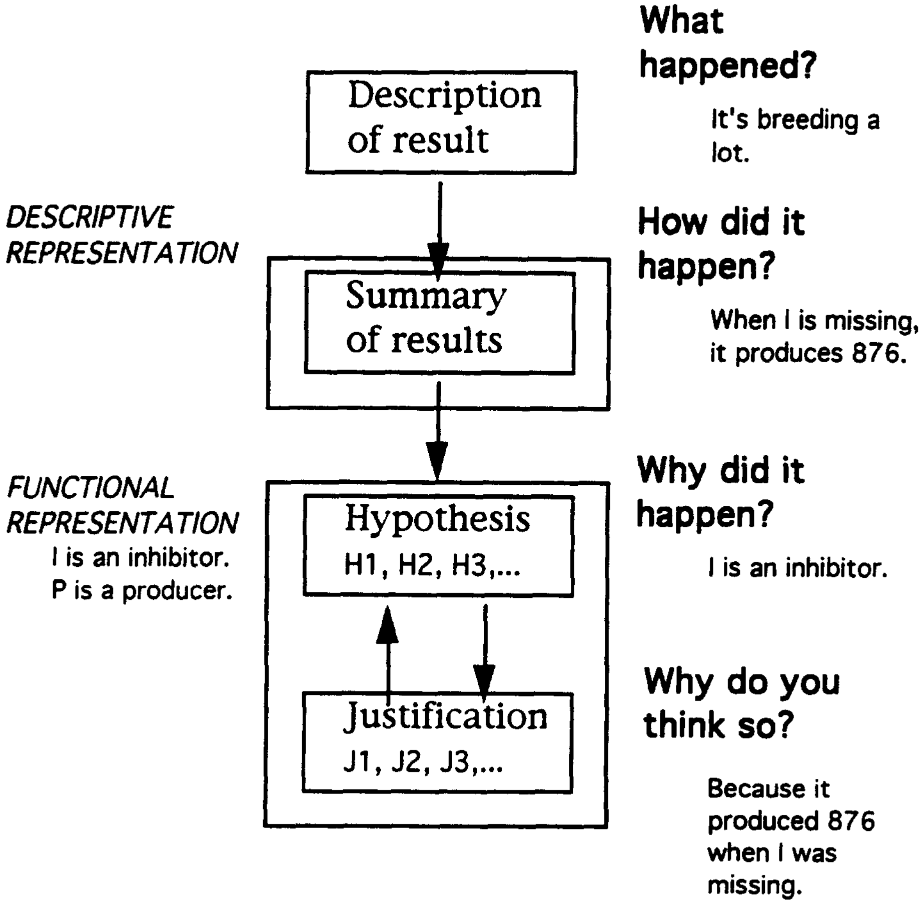


Figure 3. Level of explanation.

ized description of a result; to causal explanation of a phenomenon; to justification of the causal explanation.[4] Each level of explanation could be regarded as an answer to a specific question. For example, a description of results could answer: "What was going on?" A summary of results could answer: "How did it happen?" A causal explanation could answer: "Why did it happen?" A justification of explanation could answer: "What evidence supports the explanation?" People often ask themselves these questions, in metacognitive style: "I am not sure what it means" and "I wonder how it happened." Therefore, when we speak of requests for explanation in this paper, they include such metacognitive statements in addition to explicit questions.

In our survey of related research, we mentioned studies, of both individual and collaborative problem solving, that participation in explanatory activity leads to improved learning and performance. Not all explanations must be responses to requests. Sometimes subjects produced explanations without such requests. However, requests for explanation may indicate important decision points that require conscious, reflective thinking. They may reflect the subjects' metacognitive awareness of their cognitive processes or their decisions to initiate a new search. Whether metacognitive or not, these requests seem to prompt people to participate in explanatory activities. When subjects seek an explanation, they are standing before a door that could lead them to a new problem space or a new area of the problem space. Whether or not they can open the door depends on how they handle the request. They may not know how it can be answered and give up further search; or, they may pursue the question, trying to clarify it, and succeed in entering a new region of problem space.

In a collaborative situation, subjects must often be more explicit than in an individual learning situation, to make partners understand their ideas and to convince them. This can prompt subjects to entertain requests for explanation and construct deeper explanations. The importance of such requests on explanatory activities has received some support in previous research. Miyake (1986) showed that when people collaboratively try to understand a complex mechanical device such as a sewing machine, they deepen their understanding through an iterative cycle of understanding and non-understanding. When the subjects were in an understanding phase, their points of view were stable. Changes in point of view corresponded to periods when they did not understand. The result suggests that when people are unsure of their understanding, they tend to search actively for new points of view, and often request that their partners provide explanation. In addition to an individual's knowledge, the partners' knowledge is also available for finding answers.

Differences in availability of resources for hypothesis space search might also be an important cause for the differences in performance between the Pairs and Singles. Therefore, we tested the following two hypotheses.

Hypothesis (2-1): Pairs requested explanations more often than Singles.

This hypothesis was partially supported by the data. Table 5 shows that Pairs made more requests for explanation than Singles (38 vs. 20; $t(16) = 1.77$, $p = .10$). When the requests were divided into sub-categories, the main difference was in the requests for justification (11 vs. 1; $t(16) = 3.12$, $p < .01$). In other words, Pairs sometimes questioned whether their hypotheses were justified, while Singles rarely did so.

TABLE 5
Differences Between Paris and Singles in Terms of Reqests for Explanation

	Pairs		Singles		p of t
Measures	*Means*	*(SDs)*	*Means*	*(SDs)*	Tests
% of units with requests for explanation	38	(21)	20	(21)	=.10
with requests for description and summary of results	5	(7)	2	(3)	=.32
with requests for hypothesis	24	(22)	17	(19)	=.46
with requests for justifications	11	(8)	1	(3)	<.01
% of units with answers to requests for explanation (A/S)	80 (*n* = 9)	(19)	44 (*n* = 7)	(37)	<.05
with answers to requests for description and summary (A/S)	100 (*n* = 9)	(0)	100 (*n* = 7)	(0)	—
with answers to requests for hypothesis (A/S)	73 (*n* = 8)	(20)	48 (*n* = 7)	(43)	=.15
with answers to requests for justification (A/S)	78 (*n* = 7)	(40)	65 (*n* = 2)	(21)	

Hypothesis (2-2): Pairs answer requests for explanation more often than Singles.

This hypothesis was also partially supported. Table 5 shows that Pairs answered requests for explanation more often than Singles (80 vs. 44; $t(16) = 2.59, p < .05$). When the answers were divided into sub-categories, the main difference was found in the answers to requests for hypotheses, although it did not reach significance (73 vs. 48; $t(13) = 1.53, p = .15$).

These data suggest that requests for explanation play an important role in producing explanations. Pairs participated in such activities more often than Singles.

Collaborative Generation of Explanations

Although new explanations are generated in various situations, we pay special attention to the segments around the "requests for explanation" in the Pairs condition and provide some examples. Although our analysis is only qualitative, it offers some useful insights for further research on this issue.

We inspected all the requests for explanation coded in the previous analysis, identifying the conditions precedent to the requests and the subsequent pattern of activity. *Five types* of conditions trigger collaborators' requests for explanation (Table 6). These types include both information from the environment (i.e., experimental outcome and partner's ideas) and information from self's cognitive processes (summarizing the results, completing a sub-component, and metacognitive experience of one's own comprehension).

We also found *six types* of activity patterns that followed requests for explanation (Table 7). These patterns include both constructive activities (i.e., idea generation by a partner and a requester, data review and/or focused discussion, and postponement for information gathering) and non-constructive activities (i.e., restatement and disregard).

People respond to the requests for explanation in various ways. Requests often indicate a decision point that could lead to search a new area of a problem space or to construct a

TABLE 6
Conditions Preceding Requests for Explanation

Preceding Condition	Protocol Examples
1. Experimental outcome a puzzling new experimental result	B: (After watching the result,) What does that mean? (ss3)
2. Summarizing the results summarizing the results	A: All right. P will produce with or without lactose. I, O, without P will produce only if lactose is present. OK, so what does that mean? (ss3)
3. Completion of a sub-component grasp of a sub-component of the genetic mechanism	B: Yeah, so I and O inhibited each other. A: OK. So we have to do is to describe how it works now. (ss10)
4. Partner's ideas	
1) disagreement with partner's explanation	1) A: I is chemical and O is physical. B: Why is that? I don't think ... (ss4)
2) incomprehension of partner's explanation	2) B: So, they are chemical and physical. A: What's the difference? (ss2)
5. Metacognitive experience on one's own comprehension	
1) tenuous confidence in one's own explanation	1) A: It's turning them on and off. What do you think they are doing? (ss3)
2) sense of incomprehension	2) A: It's too complex for me. What does this result mean? (ss6)

Note: Not all of these examples are verbatim examples. Some are edited or shortened to make the examples clearer.

new representation. As Miyake (1986) pointed out, partners in a collaborative group often serve as monitors to check such decision points and to change the search to a new problem space. Although it is not always true that such activities make collaborators find correct hypotheses, they seem to have quite important roles in discovery processes.

How a Request for Explanation Changed Knowledge

An example will show how requests for explanation are generated and handled by a pair of collaborators. Near the end of their discovery process (while conducting experiment 17 and experiment 18), Pair SS8 co-constructed new knowledge based on a member's requests for explanation. The request for explanation was preceded by their summarizing the results and completing a sub-component of the task, and was followed by the partner's idea generation.

The protocol is shown in Appendix 3 and the pattern of knowledge construction is described in Figure 4, where we divide the data into six phases and describe each member's goals and knowledge in a box. We also show their interaction in the space between the two boxes. We can summarize the knowledge construction processes as follows:

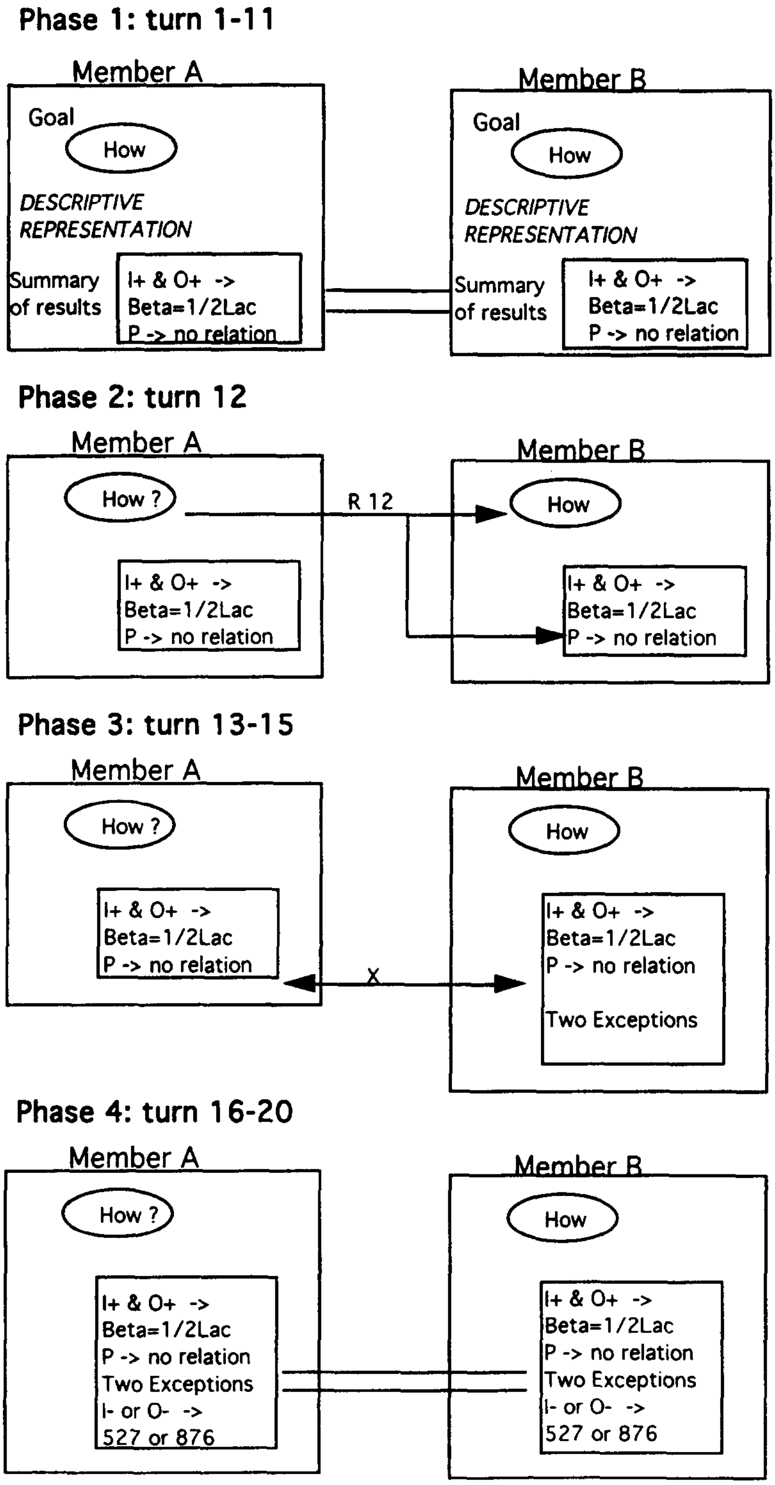
After conducting all experiments and summarizing the data, the pair reached consensus that If I+ and O+ are present, then Beta output is half of the lactose input, and P has no relation to Beta output.) (1-11). Then, after confirming their theory, A asked if they had answered

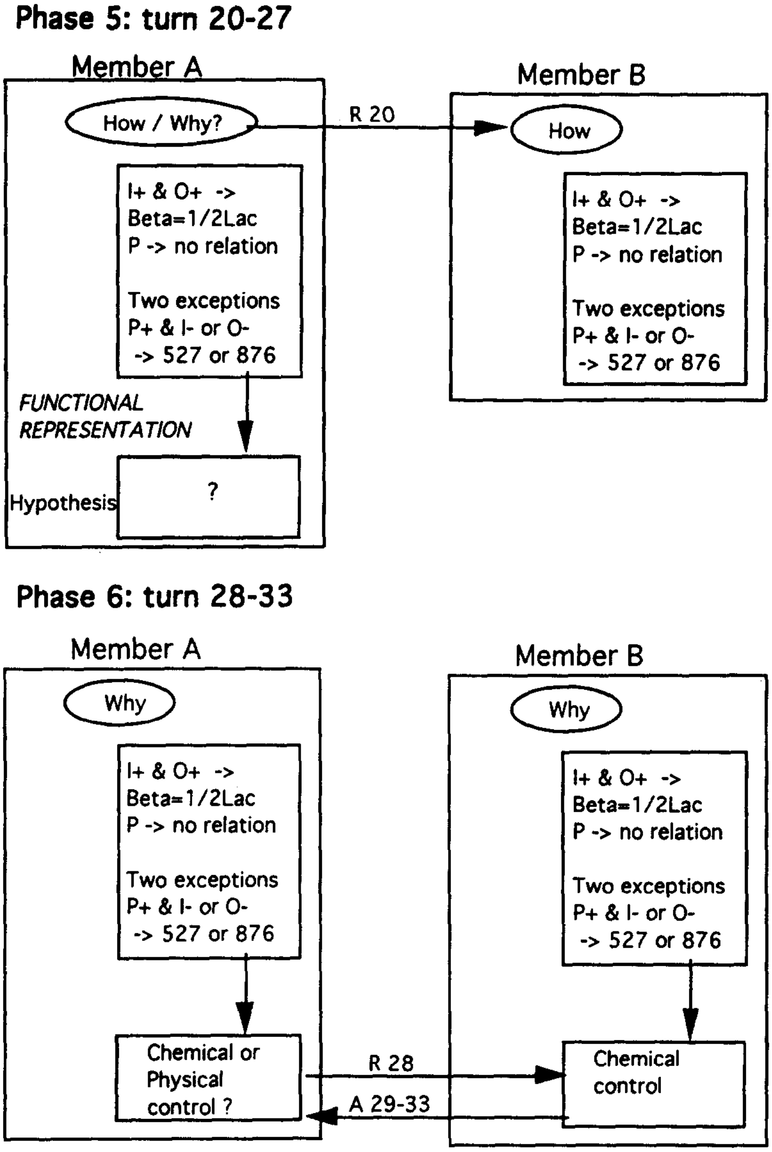
TABLE 7
Activity Pattern Following Reqests for Explanation

Activity Patterns Following Requests	Protocol Examples
1. Partner's restatement Partner simply restated a prior explanation	B: So they are chemical and physical. A: What's the difference? B. One of the genes are controlled chemically or physically. (ss2)
2. Partner's idea generation Partner offered an explanation (hypothesis and/or justification)	B: (After watching the result,) What does that mean? A. The means I shuts if off somehow ... (ss3)
3. Requester's idea generation Requester offered an explanation (hypothesis and /or justification)	B: Two Ps, one O, 527, here is two Ps, and zero Os still 527. So what does that mean? A: At least. B: it seems, ... perhaps, P is important to the reaction. (ss6)
4. Data review and/or focused discussion They reviewed data and/or discussed explanation	A: So what does that support? ... B: OK, why don't we run through, we have normal with P-, O-, and I-, they all follow the rule, right? A: In some cases, it doesn't produce any with P- ... (ss8)
5. Postponement of decision for information gathering	1) A: OK, so what can we conclude from that? B: Not requires lactose, it always produces this amount. A: What was it? 2): A: So we have to do is to describe How it works now. B: You should try the other one as well, just to make sure. Try with O, both Os are missing. (ss10)
6. Disregard of the request Partner ignored and continued his own thought	A: It's producing enzyme. B: Why is it producing here? A: The other thing is that we haven't really varied the amount of lactose. (ss12)

Note: Not all of these examples are verbatim examples. Some are edited or shortened to make the examples clearer.

the question (Request 12; 12). B pointed out two exceptions to their theory (13-15). (These could lead them to discover the chemical and physical effects). A initially denied this but agreed after a brief dispute (16-20). However, A, still feeling that his question had not been answered, stated that they were describing the "how" but not the "why" (Request 20; 20-27). This request for a "why" led the pair to discuss their goal. When A suggested that causality might be related to chemical and physical control, member B understood the partner's point and proposed that the process is chemically controlled (28-33). Although this is not a sufficient answer, it at least switched their search from a summary of results to a functional representation of processes, which is where the correct hypothesis can be found.


Phase 1: turn 1-11
Member A
Goal
How
DESCRIPTIVE REPRESENTATION
Summary of results
I+ & O+ ->
Beta=1/2Lac
P -> no relation
Member B
Goal
How
DESCRIPTIVE REPRESENTATION
Summary of results
I+ & O+ ->
Beta=1/2Lac
P -> no relation
Phase 2: turn 12
Member A
How ?
I+ & O+ ->
Beta=1/2Lac
P -> no relation
R 12
Member B
How
I+ & O+ ->
Beta=1/2Lac
P -> no relation
Phase 3: turn 13-15
Member A
How ?
I+ & O+ ->
Beta=1/2Lac
P -> no relation
X
Member B
How
I+ & O+ ->
Beta=1/2Lac
P -> no relation
Two Exceptions
Phase 4: turn 16-20
Member A
How ?
I+ & O+ ->
Beta=1/2Lac
P -> no relation
Two Exceptions
I- or O- ->
527 or 876
Member B
How
I+ & O+ ->
Beta=1/2Lac
P -> no relation
Two Exceptions
I- or O- ->
527 or 876



Note: = means that there is an agreement between two members.
<-x-> means that there is a disagreement between two members.
-R-> means that a member made a request to the partner whom the arrow points to.
-A-> means that a member answered the partner's question.
? means that the person has tenuous confidence on one's own ideas or no idea at all.

Figure 4. Pair Ss8's knowledge co-construction processes.

A's requests for explanation led them to search a new hypothesis space and to construct new knowledge. The knowledge change came about at two levels: one within a representation (12-20) and the other across representations (20-33). In both cases, requests for explanation played important roles in co-constructing the knowledge.

Subjects' Strategies

Our final analysis concerns the strategies that subjects adopted. Dunbar and Klahr (1989) found two different search strategies in a discovery task. One group of subjects "induced a correct frame from the result of an experiment in region III of the experiment space," while the other group of subjects "searched the hypothesis space for information to construct a frame that was consistent with the experimental data that they had observed" and "did not have to conduct an experiment in region III of the experiment space to generate the correct frame" (p. 124). They called the former group Experimenters and the latter group Theorists.

The concept of these different strategies is very important. We also noticed, in our protocols, that some subjects tried to search the experiment space comprehensively, taking all combinations of important variables without forming hypotheses. Other subjects entertained many hypotheses and tested the hypotheses step by step without trying many combinations of the variables.

In order to explore this distinction more fully, we focused on (a) whether subjects generated hypotheses frequently or infrequently; and (b) whether subjects considered many combinations of variables, covering a large portion of the experiment space. Using these two dimensions, we sorted the subjects into four categories. Frequency of generating hypotheses was measured by the percentage of units in which subjects entertained hypotheses (Table 3), and we divided the subjects into high and low groups relative to the mean score. Combinations of variables considered was measured by the dimension search score in Table 2. Using the mean score as the division point, we divided the subjects into high-dimension searchers and low-dimension searchers.

Subjects then have four possible strategies: (a) To generate many hypotheses and search many dimensions in the experiment space. The two subjects (Pairs) who took this strategy were termed *Comprehensive Experimenters*. (b) To generate many hypotheses without searching many dimensions in the experiment space. The seven subjects (i.e., three Singles and four Pairs) who took this strategy were called *Theory-Guided Experimenters*. (c) To generate a few hypotheses while conducting experiments that searched many dimensions. The eight subjects (i.e., five Singles and three Pairs) who took this strategy were called *Empirical Experimenters*. (d) To generate a few hypotheses without searching many dimensions in the experiment space. The one person (a Single) who fell into this category was called a *Passive Experimenter*.

The majority of the subjects (15 out of 18) were either Theory-Guided Experimenters or Empirical Experimenters. Table 8 shows the differences in experiment space search and hypothesis space search between these two strategies.

Interestingly, there were no significant differences between these two strategies in numbers of hypotheses, numbers of different types of hypotheses, types of crucial experiments,

TABLE 8
The Defferences Between Theory-Guided Experimenters and Empirical Experimenters

		Theory-Guided	Empirical	
		Experimenters	Experimenters'	
		Mean and (SD)	*Mean and (SD)*	*p* of *t*-
	Measures	(*n* = 7)	(*n* = 8)	Tests
Performance	Discovery score	2.43 (1.13)	1.88 (1.13)	=.36
Time	Time (minutes)	15.53 (3.92)	34.26 (12.31)	<.01
E-space search	Number of experiments	8.6 (2.2)	19.4 (7.0)	<.01
(Breadth)	Dimension search score	9.43 (0.79)	13.75 (1.49)	<.001*
(Informativeness)	% of types of crucial experiments	86 (10)	85 (14)	=.91
(Systemization)	Mean feature difference score	1.93 (0.23)	1.67 (0.19)	<.05
H-space search	% of units with hypotheses	83 (14)	47 (10)	<.001*
	% of units with critique	23 (27)	8 (9)	=.18
	% of units with alternative hypotheses	16 (23)	8 (9)	=.36
	% of units with combined-justification	61 (31)	33 (14)	<.05
	Number of hypotheses	20.4 (12.8)	20.3 (10.1)	=.98
	Number of types of hypotheses	8.6 (2.4)	9.1 (2.5)	=.67

Note: *These measures distinguish the two strategies by definition.

or performance scores. There were no differences in numbers of hypotheses and types of hypotheses because Empirical Experimenters spent much time and overall conducted many experiments, while Theory-Guided Experimenters talked about hypotheses intensely during a short time, conducting a small number of experiments. The difference in strategy caused a difference of efficiency in terms of time spent and the number of experiments conducted, but not in accuracy of answers.

The data suggest that there are two ways to conduct crucial experiments, depending on the strategies that subjects take. Theory-Guided Experimenters may need to have a correct hypothesis, or at least a correct question, in order to conduct all of the crucial experiments. Empirical Experimenters may need to engage in comprehensive experimentation in order to conduct all of the crucial experiments. Note that Empirical Experimenters follow the One Variable At a Time Heuristic more closely than do Theory-Guided Experimenters, whose hypotheses give additional guidance in experimental design.

We cannot affirm that one of the strategies is better than the other, for the usefulness of each strategy may depend on such conditions as subjects' background knowledge and cognitive styles. Theory-Guided Experimenters may need to participate in active thinking about the hypothesis space intensely over a short time, while Empirical Experimenters can form hypotheses more gradually when aided by systematic data collection. Therefore, if the subjects have strong background knowledge or a contemplative cognitive style that generates and tests specific hypotheses quickly, the Theory-Guided strategy may be more effective; if the subjects have neither strong background knowledge, nor a theory-oriented cognitive style, the Empirical strategy may be the more useful. This conjecture calls for further research.

GENERAL DISCUSSION

The goals of this study were (a) to compare the performances of a discovery task by Pairs and Singles, (b) to describe differences between their discovery processes, and (c) to identify the important variables that are responsible for discovery. We found that: (a) Pairs performed better than Singles; (b) Pairs participated in explanatory activities more than Singles (i.e., Pairs entertained hypotheses more often, considered alternative ideas more frequently, and talked about justification more actively); (c) Explanatory activities were effective for discovery only when the subjects also collected informative data (i.e., both conducting crucial experiments and participating in explanatory activities were necessary for discovery); (d) Explanatory activities were facilitated when subjects made requests for explanation and focused on them; and (e) Five types of conditions preceded requests for explanation and six types of activity patterns following the requests. By means of an example, we described how new knowledge was constructed collaboratively after a request for explanation. Two levels of knowledge change occurred: the one within a representation and the other across representations. Finally, two types of strategies, one emphasizing experiment space search; the other, hypothesis space search, led equally often to a solution.

On Alternative Hypotheses

Pairs entertained alternative hypotheses more often than Singles even when they were not forced to do so. Such alternative idea generation is often enhanced by a partner's requests for explanation and leads to co-construction of knowledge by collaborators.

However, we could not identify whether entertaining alternative hypotheses produced by others or entertaining alternative hypotheses produced by self is more important for discovery. Two recent papers focus on this issue (Koehler, 1994; Schunn & Klahr, 1993), using individual discovery situations to investigate whether or not other-generated hypotheses are more easily tested than self-generated hypotheses. The two papers obtained wholly opposite results. Schunn and Klahr found that the other-generated hypotheses lead to more thorough investigation of hypotheses and better performance; Koehler found that subjects who generated their own hypotheses evaluated the hypotheses more accurately than subjects who are offered other-generated hypotheses. We need further research to resolve these conflicting results and to test which effect would be found in a collaborative situation, where critique and argumentation can occur, as contrasted with an individual situation.

On Explanatory Activities

All of the recent studies, including the present one, on explanatory activities converge to tell us that these activities are a crucial component of successful intellectual behavior. Explanatory activities help people to connect pieces of information into an organized theory. Having others as monitors encourages people to participate in such activity and helps them to construct their theories more actively and more deeply (Miyake, 1986).

Many interesting questions still remained unanswered. Are there differences in explanatory activities between individual and collaborative learning situations? Do those who perform well in a problem solving task participate in the same type of explanatory activities whether they work alone or collaboratively? If we train subjects in an individual condition or a collaborative condition to participate in explanatory activities, do their problem solving processes and performance scores reach the same level? With this single study, we cannot answer these questions conclusively.

On a Computational Model for Collaborative Discovery

We are not yet ready to propose a full computational model that accounts for collaborative discovery processes. Instead, we offer some suggestions for applying the current dual space search model of scientific discovery, based on individual processes, to the collaborative discovery situation. The model offers a framework to integrate hypothesis space search and experimental space search, which, in individual problem solving, are driven by feedback from experimental outcomes or by one's own prior knowledge. The model does not address the processes that generate and change representations or the sources of discovery goals.

A theory of collaborative discovery must model two or more agents interacting. Search is driven not only by feedback from experimental outcomes or one's own prior knowledge, but also by a partner's input. A partner's questions, requests, critiques toward one's hypothesis and/or justification enhance further search in the hypothesis space and the experiment space. Also, the partner's alternative hypotheses and justifications stimulate one's hypothesis search. In this way, hypotheses and justifications are co-constructed by the members of the collaborative group. The theory must describe all agents' knowledge and goals, as well as the operations that change them.

Several different representations or goals often co-exist in a collaborative group to be negotiated and co-constructed by members. As we stated above, there are two levels of knowledge change through construction: change within a representation and change across representations. In the process of co-construction of knowledge, there are often mis-matches or conflicts between members' concepts. These may cause constructive activities, such as further search in hypothesis and experimental spaces, as well as non-constructive activities, such as disregard of the partner's requests and opinions. A model of collaborative discovery needs to include operations to handle such mis-matches and conflicts among agents. In order to do so, we need to understand the on-line processes whereby an individual understands and interprets information from his/her environment, and the environment itself changes by virtue of the individual's actions.

CONCLUSION

Scientific discoveries often are made in social situations, and due to the demands of today's society, collaborative research—including international and/or interdisciplinary collaborative research—has been emerging rapidly as the predominate form of scientific activity in many domains.

Although the importance of studying collaborative scientific discovery processes has been pointed out (Shrager & Langley, 1990), most previous studies in the psychology of science have focused on individual discovery processes. A few studies in developmental psychology (see Azmitia & Perlmutter, 1989) and group problem solving (see Hill, 1982; Levine & Resnick, 1993) have examined the details of the processes of collaboration, and these have been newly joined by some recent studies that we have discussed in our paper. Our own study identifies specific problem solving processes, notably explanatory activities and appropriate data collection, that are important to successful discovery. It also describes the ways in which these processes are accomplished and shows how they are facilitated by collaboration. It thereby takes an essential first step towards integrating studies from the psychology of science with studies from the psychology of collaboration in order to capture a broader view of scientific discovery.

APPENDIX 1
Coding Scheme of Protocol/Discourse for a Unit)

Categories	Definitions	Examples
Hypothesis (interpretation)		
Hypothesis	Hypothesis is a statement about effects of variables. When subjects mentioned a hypothesis in a unit, the unit is coded as having hypothesis. Content of hypothesis is described in Appendix 2.	The I gene chemically inhibits enzyme production.
Summary of data	Summary of data is description of data connected with condition. When subjects mentioned a summary of data in a unit, the unit is coded as having summary of data.	When the I is missing, they produce 876.
Description of result	Description of results is description of result which is not connected with its condition. When subjects mentioned a description of result in a unit, the unit is coded as having a description of result.	It's producing a lot. It produced before the lactose arrived at the chromosome.
Prediction of result	When subjects predicted the results of the next experiment in a unit, the unit is coded as having prediction of result.	It must produce 876.
Alternatives (Breadth of H-space search)		
Alternative hypothesis	When subjects mentioned two different hypotheses about a variable in a unit, the unit is coded as having alternative hypotheses.	

Critique (Disagreement) to a hypothesis	When subjects expressed disagreement to the hypothesis or mentioned an alternative hypothesis in a unit, the unit is coded as having disagreement to the hypothesis.	I don't think so.
Agreement to the hypothesis	When subjects expressed agreement to the hypothesis in a unit, the unit is coded as having agreement to the hypothesis.	Yeah, right.
Extension of the hypothesis	When subjects added a new element to the hypothesis in a unit, the unit is coded as having extension of the hypothesis.	(original hypothesis: The I is chemical.) And, the O is physical.
Justification (Depth of H-space search)		
Justification through experimental results	When a hypothesis is accompanied by summary of results to justify in a unit, the unit is coded as having justification with data. Justification with several results (next category) is included in this category.	The I is an inhibitor, since when the I is missing it produced a lot.
Justification using several experimental results	When a hypothesis is accompanied by summary of results of more than one experiment in a unit, the unit is coded as having justification with global results.	The I is an inhibitor, since when the I is missing it produced a lot, while when I is there product was normal.
Plan for new experiments to test hypotheses	When subjects planned to conduct an experiment to test a hypothesis in a unit, the unit is coded as having experiment to test hypothesis.	Let's test if the I is really an inhibitor.
Testability	If subjects talked about whether a hypothesis can be tested or how it can be tested in a unit, the unit is coded as having testability.	How can we test which of those two hypotheses is right?
Argument about justification	If subjects argued against their justification, the unit is coded as having argument about justification.	
Goal		
Suspending conclusion	When subjects mentioned suspending conclusion in a unit, the unit is coded as having suspension of conclusion.	Let's not conclude now. We need more experiments.
Requests for explanation	When subjects mentioned questions or metacognitive statement which request or imply people to generate further explanations in a unit, the unit is coded as having requests for explanation.	I don't understand what this result means. How do you interpret these data? So, what can we conclude now?

Note: A unit is a period between two adjacent experiments

APPENDIX 2
An Example of Coding Procedure for Protocol

Sample data: The unit of Experiment 7 in Pair SS 3's protocol

(A: member A; B: member B)

EXP. 7 +P-O- Lactose 0 Output 0

(Haploid mutant P experiment with 0 lactose)

1 B: See there now, it doesn't do anything.
2 A: P's got to be-
3 B: P must be the one that turns it on when lactose is present.
4 A: Ok, P negative, when P was there, hey when lactose isn't present, the output-
5 B: Wait a minute, when P wasn't present at all, wait a minute, wait a minute, maybe, well this isn't right. Because we left P out at the last time, P wasn't there, and we still know if there is no lactose present, not to produce any, maybe they need the I and O at the same time. You see what I am saying. If you don't have I and O at the same time, it might not-
6 A: Yeah, we only took out I both times, it produced-
7 B: You got O-
8 A: It produced-
9 B: So maybe when you have both of them together.
10 A: I and O turn off, turn on or turn off? But the lactose has to be present.
11 B: Try without O, and without any lactose to see what happens to make sure we cover all of the bases there. Then if-
12 A: Start reaction.

Coding

Categories	Coding	Where?
Hypothesis	Yes	3, 4, 5, 10
Summary of results	Yes	4, 5, 6
Description of result	Yes	1
Prediction of result	No	
Alternative hypotheses	Yes	3 vs.. 5 & 10
Disagreement with the hypothesis	Yes	5
Agreement to the hypothesis	Yes	4 for 3 & 6 for 5
Extension of the hypothesis	Yes	3,4 to 10
Justification through experimental results	Yes	5, 6
Justification using several experimental results	Yes	5, 6
Plan for new experiments to test hypotheses	No	
Testability	No	

APPENDIX 3
Pair SS8's discussion sample

Experiment 17: IPO- IPO- 500 / 527

(lines deleted.**Summarizing the results.**)

1 B: What we are saying is basically as long as an O and I are present, normal it will be.
2 A: That's what I am thinking.
3 B: Or, what we are saying is that if you have I+ and O+ present, right, in the thing no matter whether it's diploid or haploid, that it's going to give you, like, beta in half the amount of lactose.

(lines deleted.)

Experiment 18: I-PO IPO- 100 / 50

4 A: That should be fifty, right, they passed it out, something like that. OK, so that makes some sense, it all makes sense.

5 B: What's your last name?

6 A: Why?

7 B: Well, that will be the name of the theory. But, anyway, so, if both a normal I and a normal O...
(Writing: If both a nomal I and a normal O are present in the cell, Beta output=1/2 Lactose.)

8 A: Are present, or positive, oh present, sorry.

9 B: In the cell, all right?

10 A: No matter what that P is, no matter if P is there or not.

11 B: All right. Oh, we don't need that. That's just kinda, it can be understood.
(Completion of a subcomponent)

12 A: **Have we answered the question?**

13 B: We have two exceptions to this, right?

14 A: No, we haven't any exceptions to that.

15 B: We have two other cases we have to account for.

16 A: **Oh, why is it 527 and 876?**

17 B: Yeah, right. Let's see if-

18 A: If only I is missing then, then why-

19 B: There is no I.

20 A: So, well, we could probably figure, all right,
so then the O,
see all we are doing is describing how it, like exactly what it will do every time, we are not saying why it is.

21 B: **What's the question?**

22 A: **See in what way do the genes, in what way do they tell the genes to produce,**
see I mean we know that every time you don't have an I, either one, you know that if you either have two negative Is in the diploid, or an negative I in the haploid, you are going to get 876.
We don't know why.
We just know that that's always gonna happen.
You know what I mean?

23 B: Hm?

24 A: I mean the P is irrelevant if they are both there,
but if the P is there, and one of the other ones is missing, then you are going to get these two constants.
That doesn't make sense, I don't know why?

25 B: **OK, that's, I guess we got to figure out what happens.**

26 A: **Oh, they want us to figure out why.**

27 B: "What you should do is find out in what way..."
Well, we know in what way.

28 A: **Does this mean control chemically or physically?**

29 B: I see.
Some of the other genes that control chemically.
Oh, we have figured out that they are controlled chemically, because it doesn't matter if the I and O on the same gene or not, it's attaching or not.
Remember how, we have the same deal with the diploid and haploid even if we separated it.
(Partner's idea generation.)

30 A: OK.

31 B: This case is the most important, are the one that is I- here, and the O- there, because-

32 A: Right.

33 B: That still works normally so they are controlled chemically.

Acknowledgments: This paper is based on the Ph.D. thesis of the first author submitted to the Psychology Department of Carnegie Mellon University and was prepared for publication while the first author was working at the Learning Research and Development Center, University of Pittsburgh, as a McDonnell Foundation CSEP postdoctoral fellow. This work was supported by the A.W. Mellon Foundation under the project title: The Literacy in Science Research Center at CMU. We greatly appreciated the comments and support of Kevin Crowley, David Klahr, Jill Larkin, and Chris Schunn throughout the period of conducting this study. We thank Kevin Dunbar for allowing us to use his task. We also thank Micki Chi, Shari Ellis, Anne Fay, Ellice Forman, Jim Greeno, Buz Hunt, Jin Woo Kim, Naomi Miyake, Nancy Nersessian, Lauren Resnick, Jeff Shrager, Fritz Staub, Paul Thagard, Naoki Ueno, and one anonymous reviewer for their comments.

NOTES

1. Figure numbers in this citation were changed from Dunbar (1993) in order to match the numbers in this paper.
2. Some subjects used the word "control" instead of inhibition. In most cases, we asked the subjects to clarify the meaning. If they did not make it wholly clear but it was obvious from the context that control did not mean activation, we regarded it as inhibition.
3. Remember that Pairs' and Single's percentages of crucial experiments were nearly equal (89% vs. 87%). Hence there is no implication from the data that explanatory activities caused more crucial experiments to be generated. Instead, the hypothesis claims that explanatory activities permitted a fuller exploitation of the findings of critical experiments.
4. The upper three levels: description of results, summary of results, and causal explanation of a phenomenon are approximately the same as the notions of observation, laws, and theories in scientific knowledge in Thagard (1988).

REFERENCES

Azmitia, M., & Perlmutter, M. (1989). Social influences on children's cognition: State of the art and future directions. In H.W. Reese (Ed.), *Advances in child development and behavior* (vol. 22, 89–144). San Diego, CA: Academic.

Brown, A.L., Campione, J.C., Reeve, R.A., Ferrara, R.A., & Palincsar, A.S. (1991). Interactive learning and individual understanding: The case of reading and mathematics. In L.T. Landsmann (Ed.), *Culture, schooling, and psychological development: Human development* (vol. 4, 136–170). Greenwich, CT: Ablex.

Brown, A.L., & Palincsar, A.S. (1989). Guided, cooperative learning and individual knowledge acquisition. In L. Resnick (Ed.), *Knowing, learning, and instruction* (393–451). Hillsdale, NJ: Erlbaum.

Chi, M.T.H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

Chi, M.T.H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477.

Dunbar, K. (1989). Scientific reasoning strategies in a simulated molecular genetics environment. In *Proceedings of the Eleventh Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science, 17*, 397–434.

Dunbar, K., & Klahr, D. (1989). Developmental differences in scientific discovery processes. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 109–143). Hillsdale, NJ: Erlbaum.

Dunbar, K., & Schunn, C.D. (1990). The temporal nature of scientific discovery: The role of priming and analogy. In *Proceedings of the Twelfth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

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

Farris, H.H., & Revlin, R. (1987). *Hypothesis testing: Confirmation bias or counterfactual reasoning*. Santa Barbara: University of California at Santa Barbara.

Flor, N.V., & Hutchins, E.L. (1991). Analyzing distributed cognition in software teams: A case study of team programming during perfective software maintenance. In J. Koenemann-Belliveau, T.G. Moher, & S.P. Robertson (Eds.), *Empirical studies of programming: Fourth workshop* (pp. 36–64). Greenwich, CT: Ablex.

Freedman, E.G. (1992). Scientific induction: Individual versus group processes and multiple hypotheses. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

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

Gorman, M. E., & Gorman, M.E. (1984). A comparison of disconfirmatory, confirmatory and control strategies on Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology, 36A*, 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*, 65–79.

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive Science, 3*, 275–310.

Hill, G.W. (1982). Group versus individual performance: Are N + 1 heads better than one? *Psychological Bulletin, 91*, 517–539.

Hutchins, E.L. (1995). How a cockpit remembers its speeds. *Cognitive Science, 19*, 265–288.

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

Klahr, D., Dunbar, K., & Fay, A.L. (1990). Designing good experiments to test 'bad' hypotheses. In J. Shrager & P. Langley (Eds.), *Computational models of discovery and theory formation* (pp. 355–402). San Mateo, CA: Morgan Kaufman.

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

Klayman, J., & Ha, Y. (1989). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review, 94*, 211–228.

Koehler, D.J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology, 20*, 461–469.

Kruger, A. C. (1993). Peer collaboration: Conflict, cooperation, or both? *Social Development, 2*, 165–181.

Kruger, A.C., & Tomasello, M. (1986). Transactive discussions with peers and adults. *Developmental Psychology, 22*, 681–685.

Kuhn, D., & Phelps, E. (1982) The development of problem-solving strategies. In H.W. Reese (Ed.), *Advances in child development and behavior* (Vol. 17). New York: Academic.

Kulkarni, D., & Simon, H.A. (1988). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science, 12*, 139–175.

Langley, P., Simon, H.A., Bradshaw, G.L., & Zytkow, J.M. (1987). *Scientific discovery: Computational explorations of the creative process*. Cambridge, MA: MIT Press.

Laughlin, P.R. (1988). Collective induction: Group performance, social combination processes and mutual majority 2nd minority influence. *Journal of Personality and Social Psychology, 54*, 254–267.

Laughlin, P.R. (1991). Collective versus individual induction: Recognition of truth, rejection of error, and collective information processing. *Journal of Personality and Social Psychology, 61*, 50–67.

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., & 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., & Shippy, T.A. (1983). Collective induction. *Journal of Personality and Social Psychology, 45*, 94–100.

Levine, J.M., & Resnick, L.B. (1993). Social foundations of cognition. *Annual Review of Psychology, 44*, 585–612.

Miyake, N. (1986). Constructive interaction and the iterative process of understanding. *Cognitive Science, 10*, 151–177.

Okada, T. (1992). Multiple hypotheses and integration of hypotheses while reasoning about buoyancy forces. Paper presented at the 25th International Congress of Psychology in Brussels.

Okada, T., Schunn, C.D., Crowley, K., Oshima, J., Miwa, K., Aoki, T., & Ishida, Y. (1995). Collaborative scientific research: Analyses of historical and interview data. Paper presented at the 12th Annual Conference of the Japanese Cognitive Science Society.

Palincsar, A.S, Brown, A.L., & Martin, S.M. (1987). Peer interaction in reading comprehension instruction. *Educational Psychologist, 22*, 231–253.

Qin, Y., & Simon, H.A. (1990). Laboratory replication of scientific discovery processes. *Cognitive Science, 14*, 281–312.

Rogoff, B. (1995). Observing sociocultural activity on three planes: Participatory appropriation, guided participation, apprenticeship. In A. Alvarez, P. del Rio, & J.V. Wertsch (Eds.), *Sociocultural studies of mind* (pp. 139–164). London: Cambridge University Press.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49*, 31–57.

Schooler, J.W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General, 122*, 166–183.

Schunn, C.D., & Dunbar, K. (1996). Priming, analogy, and awareness in complex reasoning.

Schunn, C.D., & Klahr, D. (1993). Self vs. other-generated hypotheses in scientfic discovery. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 900–905). Hillsdale, NJ: Erlbaum.

Schunn, C.D., Okada, T., & Crowley, K. (1995). Is cognitive science truly interdisciplinary?: The case of interdisciplinary collaborations. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Seeger, J.A. (1983). No innate phases in group problem solving. *Academy of Management Review, 8*, 683–689.

Shrager, J., & Langley, P. (1990). Computational approaches to scientific discovery. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Siegler, R.S., & Libert, R.M. (1975). Acquisition of formal scientific reasoning by 10- and 13- year-olds: Designing a factorial experiment. *Developmental Psychology, 11*, 401–402.

Simon, H.A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L.W. Gregg (Ed.), *Knowledge and cognition* (pp. 105–128). Hillsdale, NJ: Erlbaum.

Teasley, S.D. (1995). The role of talk in children's peer collaborations. *Developmental Psychology, 31*, 207–220.

Thagard, P. (1988). *Computational philosophy of science*. Cambridge, MA: MIT Press.

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

Tukey, D.D. (1986). A philosophical and empirical analysis of subjects' modes of inquiry in Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology, 38A*, 5–33.

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

Vera A. H., & Simon, H.A. (1993). Situated action: A symbolic interpretation. *Cognitive Science, 17*, 7–48.

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