Constraints on the experimental design process in real-world science

Lisa M. Baker and **Kevin Dunbar**
McGill University Department of Psychology
1205 Dr. Penfield Avenue
Montreal, Quebec H3A 1B1 Canada
lbaker@ego.psych.mcgill.ca or dunbar@ego.psych.mcgill.ca

Abstract

The goal of the research reported in this paper is to uncover the cognitive processes involved in designing complex experiments in contemporary biology. Models of scientific reasoning often assume that the experimental design process is primarily theoretically constrained. However, designing an experiment is a very complex process in which many steps and decisions must be made even when the theory is fully specified. We uncover a number of crucial cognitive steps in experimental design by analyzing the design of an experiment at a meeting of an immunology laboratory. Based on our analysis, we argue that experimental design involves the following processes: unpacking and specifying slots in possible experimental designs, locally evaluating specific components of proposed designs, and coordinating and globally evaluating possible experimental designs. Four sets of criteria guide local and global evaluation: ensuring a robust internal structure to the experiment, optimizing the likelihood experiments will work, performing costs/benefits analyses on possible design components, and ensuring acceptance of results by the scientific community. Our analyses demonstrate that experimental design is constrained by many non-theoretical factors. In particular, the constant threat of error in experimental results lies behind many of the strategies scientists use.

Introduction

The specification of the cognitive processes underlying scientific thinking has been a central concern of cognitive theories (e.g., Dunbar, 1995, 1996; Klahr & Dunbar, 1988; Langley, Simon, Bradshaw, & Zytkow, 1987; Nersessian, 1992; Thagard, 1992). The main focus of research on scientific thinking has been on inductive reasoning and hypothesis formation; there are now many models of the cognitive processes involved in relating theory to data. However, few models address another important aspect of scientific work, the experimental design process. Most cognitive research that has addressed experimental design has focused on the relationship between theory and experiment (e.g., Klahr & Dunbar, 1988; Klayman & Ha, 1987; Kulkarni & Simon, 1988). However, there is a rich set of reasoning processes that real-world scientists use to design experiments that refers only tangentially to theory. These cognitive processes "internal" to experimental design exemplify the rich heuristics that reasoners construct amid the constraints of complex, real-world environments. The goal of this research is to present a model of the heuristics used for designing experiments in real-world science.

Following the research strategy developed by Dunbar (Dunbar, 1995, 1996), we will investigate the way that scientists design experiments by using data that we have recently collected from an immunology laboratory at a major Canadian university. We have found that one place that scientists design and modify experiments on-line is at the weekly laboratory meeting (Dunbar, 1996). Thus, the data that we will discuss in this paper consists of audio recordings of a laboratory meeting. The advantage of using laboratory meetings as our source of data is that we obtain spontaneous statements about the design of experiments by real-world scientists in a natural way, rather than using verbal protocols, or a microworld, to investigate the experimental design process.

Method

This study is part of an ongoing research project initiated by Dunbar that has investigated various aspects of reasoning and problem solving using data from real-world science laboratories (e.g., Baker & Dunbar, in preparation; Dunbar, 1995, 1996; Dunbar & Baker, 1994; Dunbar, Patel, Baker, & Dama, 1995). The data for this study was drawn from a laboratory meeting of an immunology laboratory (excerpts from the meeting are shown in Table 1). While the laboratory was guaranteed confidentiality, we were given broad access to meetings, papers, grant proposals, and other laboratory records. At meetings of this laboratory, graduate students and other laboratory members take turns presenting recent experimental results and discussing what experiments they plan to do next. Laboratory members who were present at the meeting we analyzed include the principal investigator (PI), a postdoctoral fellow (Victoria[1]), three graduate students (including Ellen and Monica), one undergraduate honors student, and two technicians. The entire meeting was tape recorded and transcribed, and then a segment of this

[1]All names and some identifying features of the experiments have been changed.

Ellen: [1] And uh I also plan to do *in situ* with the X probes, [2] and uh with each one I will try a different modification to the initial protocols s uh [3] those modifications would be aimed at increasing the accessibility of the uh probes to the target and reducing the background. [4] So that the overall aim would be to increase the signal over uh the noise ratio. [5] OK, so example of the modica uh modification that I will do is like to use fresh frozen tissues instead of uh perfused tissues . . .

PI: [6] Are you going to be able to do all this in the next two weeks?

Ellen: [7] Well I will try fresh frozen tissues, that's for sure. [8] Perfused tissues, yes. [9] Um maybe I don't know if if I will like to treat with uh chloroform. [10] But RNAase I would like to do that as well. [11] Because in one experiment in one *in situ* we can combine all these things together.

PI: [12] All this is going into one experiment?

Ellen: [13] Yes.

PI: [14] What's the tissue?

Ellen: [15] Um hamster liver and spleen . . .

PI: [16] There's a positive control in this experiment?

Ellen: [17] Positive control would be TRX cells but that [18] I will I would mainly concentrate on tissues first

PI: [19] I think at the very very least and we're really I'm talking about the bottom before we start discussing the experiment properly, [20] there ought to be TRX induced versus uninduced. [21] Which you know already gives you a signal versus no signal. [22] I think that's got to be there minimum [23] and then we can interpret from there . . .

---------- *[other aspects of experiments are discussed]* ----------

PI: [24] And amongst all the treatments that you describe . . . which one corresponds to the industry standard, [25] and how often have you tried that one?

Ellen: [26] Uh, I will I will try it for the first time the the fresh frozen tissues [27] are what they used.

PI: [28] So that's the industry standard? [29] I mean the the likeliest way to make *in situ* hybridization work on tissues?

Ellen: [30] No there are there are people using perfused tissues as well. [31] I think published. [32] Fresh frozen tissues it's harder to get [33] especially when it's spleen. [34] I think uh this treatment and [35] and then if we don't fix before it it's very hard to to get.

PI: [36] How many people are publishing that they can get RPAF signal on fixed tissue? Is it is it more than one lab?

Ellen: [37] Yes, yes, yes . . .

PI: [38] I wonder whether you shouldn't, whether you need to, invest a huge amount of effort on fresh frozen tissue that [39] now that people can tell me this uh what they think . . .

Ellen: [40] I have the tissues.

PI: [41] You have the tissues? [42] Then most of the effort is already taken care of? . . . OK . . .

Monica: [43] But if you were to get the same results from fixed and from fresh frozen [44] probably you would choose to use fixed in the future because fixed is easier for you to work with.

Ellen: [45] OK . . .

PI: [46] The impression I get is that fixed tissue [47] I mean I have my own favorites in the business OK? [48] And it's partly bias [49] and it's partly, I see their papers around. [50] But one of my favorites is the Y lab . . . [51] They work on fixed tissue [52] and they're doing a lot of stuff that interests us [53] and we would like, in fact, to be where they are. [54] So I have this personal bias that if they can do it we can do it, [55] and it looks like easier work than using fresh tissue. [56] So what I'm trying to do is to reduce the amount of work you have to do to get a result, [57] and what's frustrating me, of course, as it always does in *in situ* [58] is the length of time before we know how it's gone. [59] Any experiment most experiments end up getting done a few times before you get a good result, and many there many many failures along the way . . . [60] But I guess what you're telling me is that you're going to do experiments blindly for a while until you start reeling in the data.

Monica: [61] Even from molecular biology though you have to . . . test a couple of different ways [62] and then as soon as you hit on the one that works OK then you know go that way. [63] And that that's what it sounds like Ellen's trying to do . . . [64] Unfortunately it takes five weeks to get the answer [65] but she says that she's gonna do a number of different tests in the in the same experiment [66] and then at the end of that five weeks

PI: [67] Within the same experiment I hadn't pick up on that's where I started asking I I was afraid that we were talking about ten different experiments.

Ellen: [68] It would it would be in into two experiments . . .

Monica: [69] All right. But you could do one one week and one the next . . . And so it would be within six weeks you would you would go, [70] Oh the fixed way...works.

PI: [71] So the comparison is going to be...fresh tissue versus

Ellen: [72] Fixed . . . Yes.

Table 1: Meeting excerpts referring to possible use of fresh frozen tissue.

meeting was selected for analysis because it involved an extended discussion of possible experimental designs. Protocol analysis techniques were used to analyze comments related to the design of future experiments. These comments were analyzed to determine the general structure of the experimental design process and the particular issues researchers took into account in designing experiments.

Analyses

Summary of the Meeting

In the meeting segment analyzed, a graduate student, Ellen, and other laboratory members discussed several possible modifications Ellen might make to a particular experimental approach. We have analyzed the entire segment; however, for explanatory clarity, here and in Table 1 we will present all references to a particular problem, the issue of what type of tissue preparation method Ellen will use.

Ellen's goal at this point in her research is to get an experimental technique called *in situ* hybridization to work. *In situ* hybridization involves placing a "probe" (in this case, a DNA probe for the enzyme RPAF) on a tissue sample and observing where on the tissue the probe adheres. Ellen hypothesizes that RPAF is involved in a particular disease. If specific adherence of the probe is achieved, Ellen will be able to see which, if any, cells in the tissue are binding to RPAF. Up to the time of this meeting, Ellen has failed to get the *in situ* hybridization technique to work.

With respect to the particular issue addressed in these meeting excerpts, Ellen is considering two possible tissue preparation methods. The first, which she has used in the past, involves "fixing" the animal tissue chemically so that experiments can be done on it. The alternate method that Ellen is considering is to freeze the tissue.

At the beginning of the meeting segment analyzed (items [1-5] in Table 1; all subsequent bracketed numerals will refer to items in Table 1), Ellen is discussing experimental manipulations she may do to try to get *in situ* hybridization to work. In particular, she says [5] she may try using "fresh frozen" tissue in her next experiments instead of the fixed, perfused tissue she has used in the past.

After Ellen describes her possible experiments, the PI asks [6] which of the experimental manipulations she intends to do immediately, and Ellen responds [7] that (among other things [8-10]) she will definitely do the fresh frozen tissue manipulation. The PI next asks Ellen [14, 16] to specify other components of her proposed experiment, including what "positive control" she will use. In using the term "positive control," the PI appears to suggest that Ellen should include as a control a type of cell to which she knows binding should occur, so that she will be able to tell whether the *in situ* procedure is working properly. Using only the experimental cells, if there is no binding Ellen will not be able to tell whether her hypothesis is wrong or whether the technique is not working. At this point Ellen says that she plans [17-18] to try only the tissue manipulation with her experimental cells and not to use control cells. However, the PI argues [19-23] that she should use control cells in the experiment.

Later in the meeting, after other aspects of Ellen's experiments have been discussed, the PI tries to evaluate [24] the experiment as a whole with respect to current standards of scientific practice. Ellen claims [27] that fresh frozen tissue preparation is the standard of the field, but under questioning from the PI [28, 36] she then admits [30-31, 37] that some laboratories also use fixed tissue. She says [32-35] that the fresh frozen method is more difficult than the fixed tissue method. Monica then argues [43-44] that Ellen should favor the fixed method because it is easier. The PI also argues [46-56] that Ellen should not try to use fresh frozen tissue, since it is both harder to work with than fixed tissue and not clearly the standard of the laboratories doing this kind of work.

The ensuing discussion [57-70] is part of an effort to coordinate all the experiments Ellen wants to do. Because *in situ* hybridization requires a wait of six weeks to obtain results, she plans to do multiple experiments soon without waiting to find out the results of each one. Because of these time frame issues, Ellen will go ahead [71-72] with the fresh frozen manipulation despite laboratory members' reservations about the fresh frozen preparation method.

Analysis of the Experimental Design Process

The previous section outlined the process scientists went through in one portion of this meeting to develop a piece of an experimental design. Beginning with an experimental approach that was not working (*in situ* hybridization), Ellen and other laboratory members examined and evaluated various alterations in methods and materials that she could implement in order to get the *in situ* technique to work. The excerpts analyzed in the last section focused on one only particular issue: whether Ellen should use fresh frozen rather than fixed perfused tissue. However, a more complete analysis of the entire meeting segment indicated that processes present in this segment occurred throughout the meeting. In this section, we analyze the general experimental design process used in this laboratory. This process is depicted in Figure 1. Figure 1 does not specify a temporal order of events; rather, it portrays an overall structure that controls the experimental design process. Often, events moved temporally from left to right to bottom in Figure 1, but elements from the three different sections were intermixed throughout this meeting segment.

At the time this meeting occurred, Ellen had already chosen to the use the *in situ* hybridization experimental approach or paradigm. Experimental paradigms are rarely constructed from scratch; rather, they may be retrieved from experiments previously done by the particular scientist, in the same laboratory, or in other laboratories (cf. Dunbar, 1996). The choice of a particular experimental paradigm constitutes the creation of a frame with particular slots corresponding to features in the experimental design (cf. Friedland & Iwasaki, 1985; Schunn & Klahr, 1995).

Unpack/specify slots in experimental design. The first part of the experimental design process exhibited by the scientists was "unpacking" the design and specifying component elements of the experiment. That is, at the

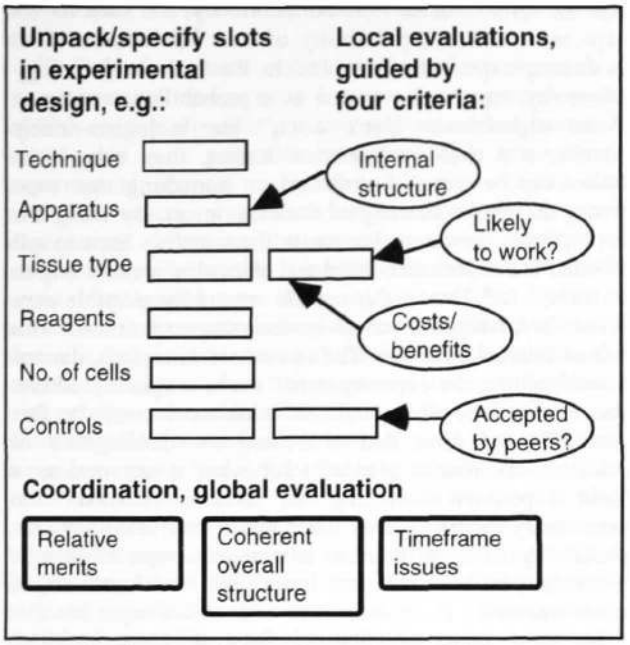

Figure 1: Diagram of general experimental design process.

beginning of the meeting Ellen had specified that she would continue to use the *in situ* hybridization technique and that she would attempt the fresh frozen method of tissue preparation. However, there were still many elements of the experimental design that had not been specified, and the PI and other laboratory members prompted Ellen to specify these [14, 16]. The scientists are able to unpack the experimental design in this way because they know the "slots" in the experimental design that need to be filled for any given experimental approach.

Local evaluation of design elements. Subsequent to, or simultaneous with, the unpacking and specification process, the scientists often evaluated specific elements of the proposed design. This type of evaluation, which was constrained to only one or a few elements of the proposed design, may be thought of as "local evaluation." Evaluation is guided by at least four sets of criteria. First, the scientists attempt to design experiments that will be "interpretable;" that is, that have a robust internal structure. This is often achieved through the use of control conditions (e.g., [16-23]) that enable the scientist to interpret the meaning of results on non-control, or experimental, conditions. Second, the scientists try to optimize the chances that the experiment will "work;" that is, that results will not be thrown off by experimental error. In this meeting the scientists tried to make sure the experiment would work by evaluating Ellen's level of expertise with the fresh frozen tissue preparation method [25] and using methods that had a proven track record (i.e., had been used successfully in other laboratories [24, 28, 36, 46-54]). Third, the scientists performed what might be considered a costs/benefits analysis. For instance, they considered how easy or difficult different methods would be to implement [32-35, 40-44, 55-56]. Fourth and finally, the scientists tried to design experiments whose results would be accepted as valid by the larger scientific community. For instance, elsewhere in this meeting the PI urged Ellen to include a particular probe because it was "an industry standard of something that you ought to find . . . everybody agrees it ought to be there."

Coordination and global evaluation. The final step in the experimental design process is coordination of different possible experiments and a global evaluation of how different approaches fit together and which experiments should be done in what order (e.g., [57-70]). Global evaluations are still guided by the same criteria described in the last paragraph, but they are used to judge the merits of different design elements and approaches relative to each other.

Discussion

The findings reported in the previous section indicate that scientists have well-developed strategies for designing and choosing among possible experiments. The analysis of these strategies suggests interesting theoretical interpretations, some of which are explored in this section.

Experimental Design: Beyond the Hypothesis

As we argued in the introduction, experimental design has most often been portrayed in the cognitive literature as a process of instantiating variables of a hypothesis in the features of an experiment. We do not deny that the scientists' current hypotheses constrain the type of experiments they choose. However, we argue that hypotheses do not *completely* constrain the design of experiments. That is, there are many possible experiments that could be used to test any given hypothesis. Even after the hypothesis is determined, the scientist still has to choose among experimental techniques, protocol steps, control conditions, and many other elements before the experimental design is complete.

In the example analyzed for this case study, Ellen had formed her hypothesis many months before the meeting, and yet she was continuing to struggle with the experimental component of her work. Ellen's hypothesis, in fact, is only obliquely referred to in the entire text of this meeting. In part, this may be because she has already chosen the experimental paradigm she will use (*in situ* hybridization); choice of experimental paradigm is perhaps more likely to involve reference to theory than later steps in the design process (cf. Friedland & Iwasaki, 1985).

Instead of translating hypothesis to experiment, the laboratory members focus on solving a technical problem (getting a new technique to work) and how to design experiments using this technique that will give interpretable results. The four sets of criteria that guide evaluation have little to do with the constraints imposed by the scientist's hypothesis. That is, a good experiment will not only test the scientist's theory, but will also be robust in internal structure (i.e. the results are expected to be interpretable), will involve methodologies at which are likely to work, will be optimal from a costs/benefits standpoint, and will produce results that will be accepted by the larger scientific community. It is particularly interesting to note that there

are complex reasoning processes involved in experimental design even after the process of translating from theory to experiment has been completed. Experimental design has its own "internal" heuristics and reasoning processes separate from the task of relating hypothesis to experiment.

There is a limited amount of earlier research that considered elements of the design process independent of theory-experiment coordination. Our results correspond to this earlier work in some ways. The interpretability criteria highlighted in our analysis correspond to the notion of "observability" of experiments developed in Klahr, Dunbar, and Fay (1990). In both cases, the scientists and subjects tried to design experiments whose results could be interpreted unambiguously. The general structure of our model of experimental design, in which slots in a frame are unpacked and specified, is similar to Friedland and Iwasaki's (1985) model of experimental design as refinement of skeletal plans. Therefore, our work is consistent with earlier research that investigated aspects of experimental design other than theory-to-experiment translation. This research may be seen as the beginning of a body of work that goes beyond the hypothesis and investigates the complex processes "internal" to the experimental design task.

Effect of Potential Error on Experimental Design

In most cognitive models of science, including psychology laboratory experiments, artificial intelligence models, and normative philosophical models, the subject/scientist runs one experiment (with one experimental condition and no control conditions) and obtains one outcome, which is presumed to be correct. In other words, experimentation in most cognitive models appears unproblematic compared to the elaborate processes we have uncovered in real-world science, where scientists struggle to choose between alternate methods and materials and design multi-condition experiments with numerous controls. The question we address in this section is: Why is the real-world process so much more complex than that depicted in earlier cognitive models?

We argue that the issue of "potential error" is what underlies many of the criteria real-world scientists take into account when designing experiments. It is a commonplace among practicing biologists that at least as many experiments "don't work" as "do work" (cf. [59]). Techniques that appear straightforward in the biology textbook or the equipment manual are in practice difficult to implement. Among other potential problems, materials may be contaminated, cells may die if not maintained in precise growing conditions, reagents may fail to adhere to or interact with other materials, and/or equipment may not be calibrated properly or may malfunction. Scientists in the laboratories studied often spent weeks or months attempting to "get a system working," and even then degradation of materials or problems implementing protocols could cause some experiments not to work.

The issue of error in experimental results has been addressed before in the psychological literature (e.g. Gorman, 1986). However, potential error has generally been treated as a probabilistic issue in the psychology laboratory; that is, on any given experiment there is some percentage chance that the result will be reported incorrectly, and subjects are expected to take this possibility of error into account when evaluating experimental results. In the real-world biology laboratory, error is not treated as a probabilistic construct. When experiments "don't work," the biologists rarely consider it a chance occurrence. Rather, they believe the failure can be causally attributed to something that went wrong during the running of the experiment. By using the appropriate experimental controls, the scientists hope to tell whether the experiment failed and also what went wrong to cause it to fail. Hence, this need to control for possible error drives the strategies involved in designing experiments with robust internal structure. The second set of criteria, directed toward getting the experiment to "work," explicitly address this issue of possible error. In addition, it may be this possibility of error that is behind the development of scientific community standards for what is accepted as a valid experimental finding: by setting standards, the community ensures that scientists can evaluate the possibility of error in other laboratories' reported results. Thus the possibility of error factors into the fourth set of criteria as well.

In short, what was unproblematic in most cognitive models—trusting the results of experiments—becomes highly problematic in the real-world science laboratory. The criteria scientists take into account when designing and evaluating potential experimental designs can be seen as a well-developed set of heuristics that have arisen to deal with the constraints of this task environment, in particular the potential for error in experimental results.

Toward a General Model of Experimental Design

Researchers building models of human problem solving have long recognized that the task environment severely constrains the types of strategies that the problem-solver will employ (e.g., Newell & Simon, 1972). Experimental design is no exception: Designing an experiment in a science laboratory is a complex problem with multiple constraints. The experimental design process depicted in Figure 1 is in fact a collection of strategies used by scientists in one laboratory to solve a portion of the experimental design problem. To the extent that these strategies have been developed by scientists in response to a task environment, it is probable that similar strategies will be manifested in laboratories confronted with similar task environments. In other words, the process depicted in Figure 1 will likely lead toward a general model of the experimental design process throughout science if laboratories in other science disciplines face similar constraints to those faced in this immunology laboratory.

We argue that the task environment for experimental design in most science domains is more similar than different from discipline to discipline and from laboratory to laboratory. For example, the particular slots scientists must fill in to complete an experimental design will vary dramatically from discipline to discipline, but the strategy of "unpacking" the design so as to make particular components of it available for analysis is likely to be used in science laboratories of many disciplines, because the structure of

experiments in modern scientific disciplines is almost always very complex.

Similarly, the criteria used to locally evaluate elements of an experimental design are a response to the general nature of 20th-century scientific practice, and not ad hoc heuristics useful only in a particular laboratory. The standards of experimental science as it currently exists require that scientists ensure a robust internal structure to their experiments. Likewise, the extensive social structure of science today requires that scientists keep in mind the likely community response to their reported results. Particular issues within these categories may vary in different scientific domains, but practitioners in all modern science laboratories must take into account these general criteria. Similarly, cutting edge science in all disciplines requires the use of new, imperfect techniques; indeed, it is the use of these techniques that often defines "cutting edge" science. Thus costs/benefits decisions and concerns about individual expertise are always present, and, more generally, when each experiment represents a large commitment of time, effort, and money, there is a need to prioritize and coordinate different experiments. For these reasons, it is not unlikely that the basic features of the model presented in Figure 1 would be present in any general model of real-world experimental design.

Acknowledgements

This research was supported by a McGill Major Graduate Fellowship to the first author. The second author was funded by grant number OGP0037356 from NSERC and grants from FCAR and SSHRC.

References

Baker, L. M. & Dunbar, K. (in preparation). How scientists deal with unexpected findings: A discovery heuristic.

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

Dunbar, K. (1996). How scientists think: Online creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Conceptual structures and processes: Emergence, discovery, and change.* Washington, DC: American Psychological Association Press.

Dunbar, K., & Baker, L. (1994). Goals, analogy, and the social constraints of scientific discovery. *Behavioral and Brain Sciences, 17,* 538-39.

Dunbar, K., Patel, V., Baker, L., & Dama, M. (1995, November). *Group reasoning strategies in knowledge-rich domains.* 36th Annual Meeting of the Psychonomic Society, Los Angeles.

Friedland, P. E., & Iwasaki, Y. (1985). The concept and implementation of skeletal plans. *Journal of Automated Reasoning, 1,* 161–208.

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.

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 scientific discovery and theory formation.* San Mateo, CA: Morgan Kaufmann.

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

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 processes.* Cambridge, MA: MIT Press.

Nersessian, N. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In R. N. Giere (Ed.), *Minnesota studies in the philosophy of science, Vol. XV: Cognitive models of science.* Minneapolis: University of Minnesota Press.

Newell, A., & Simon, H. A. (1972). *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

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

Thagard, P. (1992). *Conceptual Revolutions.* Cambridge, MA: MIT Press.