That s Odd! How Scientists Respond to Anomalous Data

Susan B. Trickett
(stricket@gmu.edu)
Dept. of Psychology
George Mason University
Fairfax, VA 22030 USA

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

Christian D. Schunn
(schunn@gmu.edu)
Dept of Psychology
George Mason University
Fairfax, VA 22030 USA

Anthony Harrison
(aharris8@gmu.edu)
Dept. of Psychology
George Mason University
Fairfax, VA 22030 USA

Abstract

We use an *in vivo* methodology to investigate the responses of scientists to anomalies. Protocols of 3 scientists performing data analysis in 2 domains were analyzed. We found that the scientists noticed anomalies and paid more attention to them than to expected data. This attention took the form of proposing a hypothesis and then elaborating that hypothesis by reference to other data in the visual display, rather than to the scientists theoretical domain knowledge.

Introduction

How do scientists deal with anomalous or unexpected data? Philosophers of science (e.g., Kuhn, 1962; Lakatos, 1976) have argued that anomalies play a crucial role in moving science forward. Scientists themselves have claimed that investigating anomalies lies at the heart of scientific innovation (e.g., Knorr, 1980).

Psychologists have been interested in the cognition underlying how scientists deal with anomalies. There have been two general approaches to the study of anomalies. One approach focuses on response to negative evidence in concept identification tasks (Wason, 1960). Several studies have found that people are likely to seek confirming evidence for their theories (e.g., Mynatt, Doherty, & Tweney, 1977).

Surprisingly, studies from this tradition have found that scientists are also very susceptible to confirmation bias (e.g., Mahoney & DeMonbreun, 1977). One criticism of this approach is that the tasks are abstractions of the hypothesis-testing cycle and therefore do not allow the participants to make use of their extensive domain knowledge (e.g., Chinn & Malhotra, 2001). However, sociological studies (e.g., interviews) of practicing scientists have also found that scientists appear to display confirmation bias (Mitroff, 1974).

A second approach investigates scientists' response to anomalies as they perform analyses of authentic scientific data. This approach includes both historical studies of scientific discovery (Chinn & Brewer, 1992; Kulkarni & Simon, 1988; Nersessian, 1999) and *in vivo* observations of contemporary practicing scientists (Dunbar, 1997; Trickett, Trafton, & Schunn, 2000). Chinn and Brewer have developed a taxonomy of responses, from ignoring the anomaly to changing the theory. They have found evidence for the whole range of responses in historical records of science. Yet the results of other studies suggest that scientists do pay attention to anomalies. For example, Dunbar (1997) found that individual scientists were quick to discard a hypothesis when faced with results that were inconsistent with it, and Kulkarni and Simon (1988) identified an "attend to surprising result" heuristic as crucial to Hans Krebs' discovery of the urea cycle.

Recently, Alberdi, Sleeman and Korpi (2000) have brought together these two approaches to the study of anomalies in scientific thinking. They conducted a psychological study of expert botanists performing a categorization task in the domain of plant taxonomy. As participants formed hypotheses about the category into which a current set of plants would fall, they were presented with a "rogue," or anomalous, item that belied their expectations. Alberdi and his colleagues found that participants did indeed pay attention to the anomalies. Furthermore, they identified a number of strategies by which participants attempted to resolve the anomalous data. Key among these was the "instantiate" strategy, in which participants searched their theoretical domain knowledge for a new hypothesis that would accommodate the anomaly.

Although categorization is an important task in many areas of science, there are many other situations in which scientists might encounter anomalies. For example, to name a few, experimental results might fail to match a prediction, a computational model might yield a different output from expectation, or empirical data might contain puzzling phenomena hard to explain by means of current theoretical understanding. Thus, many questions remain about other circumstances under which a scientist encounters an anomaly in the course of his or her own research.

The goal of this paper is to investigate scientists' response to anomalies they discover as they analyze their own data. Our first question concerns the extent to which scientists attend to anomalous data. One can imagine a range of possible responses, from ignoring the anomaly to attempting to give a full accounting of it (Chinn and Brewer, 1992). Do practicing scientists tend to ignore anomalies as suggested by Mitroff (1974) or do they attend to anomalies as suggested by Dunbar (1997) and Alberdi et al. (2000)? Our second question concerns the processes and strategies by which scien-

tists deal with anomalies when they encounter them. Do they take a theoretical approach, as found by Alberdi et al (2000) or focus on the data itself? Alternatively, additional strategies might emerge from observing scientists at work. We investigate all these possibilities.

Method

In order to investigate the issues discussed above, we have adapted Dunbar s in vivo methodology (Dunbar, 1997; Trickett, Trafton & Schunn, 2000). This approach offers several advantages. It allows observation of experts, who can use their domain knowledge to guide their strategy selection. It also allows the collection of "on-line" measures of thinking, so that the scientists' thought processes can be examined as they occur. Finally, the tasks the scientists do are fully authentic.

Two sets of scientists were videotaped while conducting their own research. All the scientists were experts, having earned their Ph.D.s more than 6 years previously. In the first set, two astronomers, one a tenured professor at a university, the other a fellow at a research institute, worked collaboratively to investigate computer-generated visual representations of a new set of observational data. At the time of this study, one astronomer had approximately 20 publications in this general area, and the other approximately 10. The astronomers have been collaborating for some years, although they do not frequently work at the same computer screen and the same time to examine data.

In the second dataset, a physicist with expertise in computational fluid dynamics worked alone to inspect the results of a computational model he had built and run. He works as a research scientist at a major U.S. scientific research facility and had earned his Ph.D. 23 years ago. He had inspected the data earlier but made some adjustments to the physics parameters underlying the model and was therefore revisiting the data.

Both sets of scientists were instructed to carry out their work as though no camera were present and without explanation to the experimenter (Ericsson & Simon, 1993). The relevant part of the astronomy session lasted about 53 minutes, and the physics session, 15 minutes. All utterances were later transcribed and segmented according to complete thought. All segments were coded by 2 coders as on-task (pertaining to data analysis) or off-task (e.g., jokes, phone interruptions, etc.). Inter-rater reliability for this coding was more than 95%. Off-task segments were excluded from further analysis. On-task segments (N = 649 for astronomy and N = 176 for physics) were then grouped into episodes (N = 19 for astronomy and N = 9 for physics). Episodes began with the scientists' focus on a phenomenon and lasted until attention switched to another feature or theoretical issue. This grouping of the protocol into episodes allowed us to focus on the more immediate reaction to anomalies.

The Tasks and the Data

Astronomy The data under analysis were optical and radio data of a ring galaxy. The astronomers high-level goal was to understand its evolution and structure by understanding the flow of gas in the galaxy. In order to understand the gas flow, the astronomers must make inferences about the velocity field, represented by contour lines on the 2-dimensional display.

The astronomers task was made difficult by two characteristics of their data. First, the data were one- or at best 2-dimensional, whereas the structure they were attempting to understand was 3-dimensional. Second, the data were noisy, with no easy way to separate noise from real phenomena. Figure 1 shows a screen snapshot of the type of data the astronomers were examining. In order to make their inferences, the astronomers used different types of image, representing different phenomena (e.g., different forms of gas), which contain different information about the structure and dynamics of the galaxy. In addition, they could choose from images created by different processing algorithms, each with advantages and disadvantages (e.g., more or less resolution). Finally, they could adjust some features of the display, such as contrast or false color.

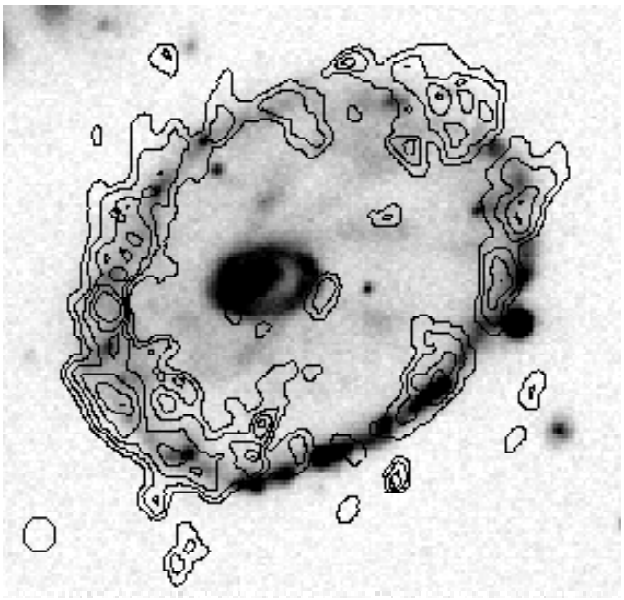


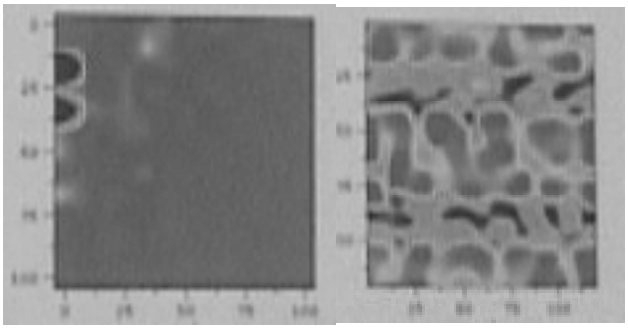
Figure 1: Example of data examined by astronomers. Radio data (contour lines) are laid over optical data.

Physics The physicist was working to evaluate how deep into a pellet a laser light will go before being reflected. His high-level goal was to understand the fundamental physics underlying the reaction, an understanding that hinged on comprehending the relative importance and growth rates of different modes. The physicist had built a model of the reaction; other scientists had independently conducted experiments in which lasers were fired at pellets and the reactions recorded. A

close match between model and empirical data would indicate a good understanding of the underlying theory. Although the physicist had been in conversation with the experimentalist, he had not viewed the empirical data, and in this session he was investigating only the results of his computational model. However, he believed the model to be correct (i.e., he had strong expectations about what he would see), and in this sense, this session may be considered confirmatory.

The data consisted of two different kinds of representation of the different modes, shown over time (nanoseconds). The physicist was able to view either a Fourier decomposition of the modes or a representation of the raw data. Figure 2 shows an example of the physicist's data. He could choose from black-and-white or a variety of color representations, and could adjust the scales of the displayed image, as well as some other features. He was able to open numerous views simultaneously. A large part of his task was comparing images, both different types of representation of the same data and different time slices represented in the same way.

Figure 2: Example of data examined by physicist
Fourier modes (left) and raw data (right)



Coding Scheme

Our goal in this research was to investigate scientists' response to anomalous data. First, we wanted to establish whether and to what extent the scientists noticed and attended to anomalies. Second, we wanted to investigate the processes by which they respond.

Both protocols were coded independently by 2 different coders. Initial inter-rater reliability for each code was greater than 85%. Disagreements were resolved by discussion. Any coding disagreements that could not be resolved were excluded from further analysis.

Noticings In order to establish which phenomena unusual or not the scientists attended to, we first coded for the scientists' *noticing* phenomena in the data. A noticing could involve merely some surface feature of the display, such as a line, shape, or color, or it could involve some interpretation, for example, identifying an area of star formation or the implied presence of a mode. Only the first reference to a phenomenon was coded as a noticing; coding of subsequent references to the same phenomenon is discussed below.

Because our investigation focused on the extent to which the scientists attended to anomalies in the data, we further coded these noticings as either "anomalous" or "expected," according to one or more of the following criteria: a) in some cases the scientist made explicit verbal reference to the fact that something was anomalous or expected; b) if there was no explicit reference, domain knowledge was used to determine whether a noticing was anomalous or not;[1] c) a phenomenon might be associated with (i.e., identified as like) another phenomenon that had already been established as anomalous or not; d) a phenomenon might be contrasted with (i.e., identified as unlike) a phenomenon that had already been established as anomalous or not; e) a scientist might question a feature, thus implying that it is unexpected. Table 1 illustrates these codes.

Criterion	Code	Example
Explicit	Anomalous	What's *that funky thing* That's odd
Domain Knowledge	Expected	You can see that *all the H1* is concentrated in the ring
Association	Anomalous	You see *similar kinds of intrusions* along here
Contrast	Expected	That's odd As opposed to *these things*, which are just the lower contours down here
Question	Anomalous	I still wonder why *we don't see any H1 up here* in this sort of northern ring segment?

Table 1: Noticings (in italics): anomalous or expected

Subsequent References One of our questions was the extent to which the scientists attended to anomalies. The coding of noticings captured only the first reference to a phenomenon of interest; we needed to establish how frequently they made subsequent reference to each noticing. Consequently, all subsequent references were also identified and coded.[2] Not all subsequent references immediately followed a noticing; frequently, the scientists returned to a phenomenon after investigating other features of the data. Subsequent references were identified both within the episode in which the noticing had occurred and across later episodes.

The rest of the coding scheme addresses *how* the scientists responded to the anomalies, in particular imme-

[1] The coders domain knowledge came from textbooks and interviews with the scientists.

[2] In the astronomy dataset, because the scientists shared a computer monitor, frequently the first interaction between them after a noticing was to make sure they were both looking at the same thing. Subsequent references that served purely to establish identity were *not* included in the analyses.

diately after they notice the anomalies. To investigate the scientists immediate response to their anomalous findings, we coded 10 utterances following each noticing, whether anomalous or expected (minus utterances establishing which phenomenon was under discussion, in the astronomy dataset). We anticipated that scientists would attempt to produce hypotheses for the anomalies, and that some of these hypotheses would be discussed further. Based on the results reported by Alberdi, et al. (2000), we investigated the extent to which elaboration of hypotheses was grounded in theory or in the visual display of the data. We also anticipated the use of additional strategies and inspected the data to identify strategies that emerged, as discussed below.

Hypotheses Statements that attempted to provide a possible explanation for the data were coded as hypotheses. All hypotheses were further coded as *elaborated* or *unelaborated*. Elaboration consisted of one or more statements that either supported or opposed the hypothesis. Hypotheses that were not discussed after they were proposed were coded as unelaborated.

When a hypothesis was elaborated, we coded whether the elaboration was *theoretical* or *visual*. When evidence for or against a hypothesis was grounded in theoretical domain knowledge, elaboration was coded as theoretical; when evidence came from the display, it was coded as visual.

Place in context A strategy that emerged from our examination of the data was considering the noticing in relation to other data. Thus we coded whether or not the scientist placed the noticing in context, and whether that context was another part of the dataset (*local*) or the scientist's own theoretical knowledge (*global*).

Results and Discussion

Noticing Anomalies

Our first question was did the scientists notice anomalies in the data?[3] Recall that a noticing is a first-time reference to a phenomenon of interest. Table 2 presents the total number of noticings for each dataset and the percentages of anomalous and expected phenomena. As Table 2 shows, at least one-third of the phenomena the astronomers identified and almost one-half the physicist identified were unusual in some way. It appears then that the scientists *did* notice anomalies in their data.

	Total Noticing	Anomalous	Expected	Not coded
Astronomy	27	33%	48%	19%
Physics	9	44%	44%	12%

Table 2: Frequency of anomalies and expected noticings

[3] We presented a more detailed discussion of a subset of the results for the astronomy dataset in Trickett et al. (2000).

Attention to Anomalies

Once the scientists had identified something unusual in the data, what did they do with this observation? There are several possible reactions: they could pursue the anomaly in order to try to account for it, they might temporarily disregard it but return to it later, or they might move on to explore some other, better understood, aspect of the data. A related question is whether their response to anomalies was different from their response to expected phenomena.

We investigated this issue by counting how often the scientists made subsequent reference to a noticing immediately upon identifying it. If anomalies and expected phenomena are of equal interest, we would expect them to make a similar number of references to both the anomalous and expected patterns. However, if anomalies play a more important role in their efforts to understand the data, we would expect them to pay more attention (measured by the number of subsequent references) to anomalies than to expected observations.

As Table 3 shows, for both the astronomy and physics datasets, scientists paid more attention to anomalies than expected phenomena, $t(28)=2.33$, $p<.05$. In the case of astronomy, the anomalies received over 3 times as many subsequent references within the same episode as the expected phenomena. The physics dataset follows a similar pattern, with more than twice as many references to anomalies as expected phenomena. The results are in stark contrast to the findings of the confirmation bias literature.

	Anomalies	Expected
Astronomy	7.6	1.5
Physics	3.0	1.25

Table 3: Mean number of subsequent references per noticed object to anomalies and expected phenomena

Immediate Response to Anomalies

We have shown that when the scientists noticed an anomaly, they immediately attended to it, but we have not analyzed the content of that attention to anomalies. In order to understand what kind of the scientists made, we now turn to the results of the second part of our coding scheme, which was applied to the 10 utterances that immediately followed the initial noticing of anomalies and expected phenomena.

Identify Features As Figure 3 shows, the scientists were only slightly (and nonsignificantly) more likely to identify specific features of the anomalies as the expected noticings, and this pattern held for both domains.

Propose Hypothesis As Figure 4 shows, the scientists were much more likely to propose a hypothesis for the

anomalies than the expected noticings $\chi^2(1) = 7.5$, $p<.05$, and this pattern was very strong in both domains.

Figure 3: Percentage of noticings for which scientists identified features.

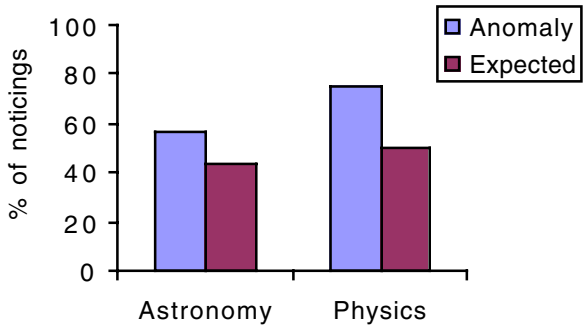
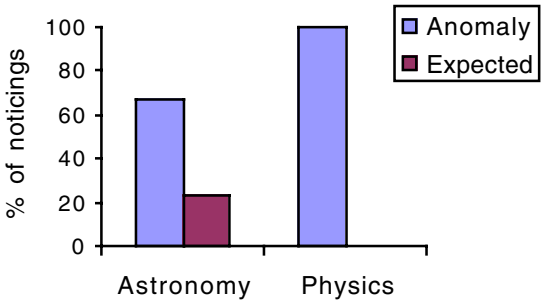
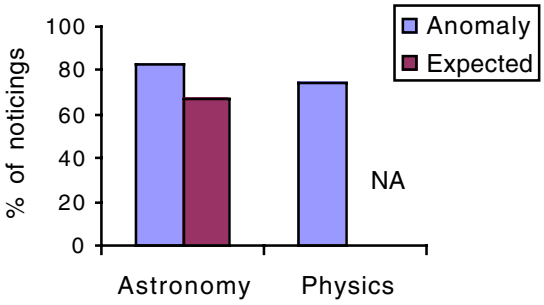


Figure 4: Percentage of noticings with hypotheses



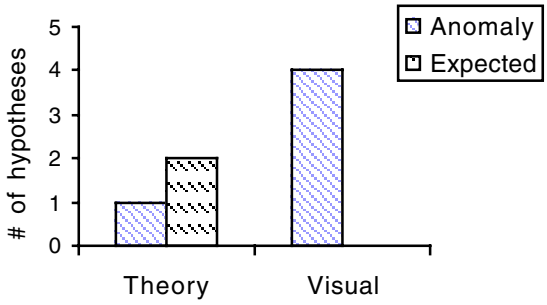
Elaborated Hypothesis Once the scientists had proposed a hypothesis (primarily about the anomalies), in most cases they elaborated that hypothesis. Figure 5 presents the proportion of hypotheses that were elaborated within each domain for expected and anomalous noticings. In most cases, scientists attempted to elaborate the hypotheses, for both expected and anomalous noticings (note that there were no hypotheses to elaborate in the expected physics case).

Figure 5: Percentage of noticings that were elaborated



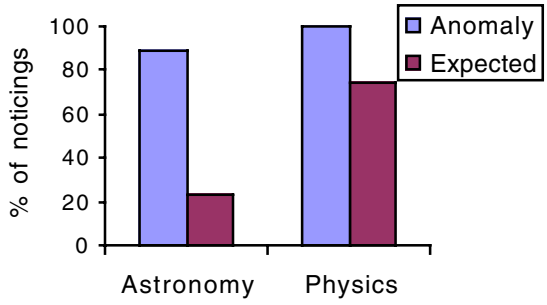
Source of Elaboration For the physics dataset, there were not enough elaborated hypotheses to analyze further. For the astronomy data, evidence about 4 of the 5 hypotheses about anomalies came from the visual display. The 2 hypotheses about expected noticings were developed theoretically. Figure 4 shows this result.

Figure 6: Elaboration type for hypotheses (astronomy)



Place in Context In addition to (or instead of) developing hypotheses about the noticings, the scientists also might consider the noticing in relation to other information, either theoretical information in memory (global context) or information about the current dataset (local context), or they might not place it in either context. In fact, none of the noticings was considered in the context of the scientists' theoretical knowledge (global). However, the scientists considered the noticings in the context of the current dataset (local), and this sequence occurred more frequently for the anomalies than for the expected phenomena, especially in the astronomy dataset (see Figure 7), $\chi^2(1) = 9.21$, $p<.01$.

Figure 7: Percentage of noticings put in local context



General Discussion and Conclusion

We examined the behavior of scientists at work, analyzing their own data. Our results show that these scientists not only notice anomalies in the data, but also attend to them, contrary to the confirmation bias literature, but similar to the findings of Dunbar (1997) and Alberdi et al. (2000).

The scientists we observed not only notice and attend to anomalies, but also do so in a particular way. Furthermore, this pattern is quite different from the pattern that results from their observation of expected phenomena. When they notice an expected phenomenon, after identifying or describing its features, the scientists are likely to engage in no further elaboration of the phenomenon. On the rare occasions when they do attempt to account for it by proposing a hypothesis, they seek evidence in their own theoretical knowledge, rather than in the visually displayed data. By contrast, however, for anomalous noticings, the scientists attempt to account for the anomaly by proposing a hypothesis. They then elaborate the hypothesis, primarily by seeking evidence in the visual display, and finally consider how the anomaly relates to neighboring phenomena within the same display.

Our results mesh in part with those of other researcher, in that they provide further evidence for the important role played by anomalies as scientists analyze and reason about data. However, our results differ from those of Alberdi et al. (2000) in some significant ways. When the botanists in their study encountered an anomaly, they were most likely to use a strategy of theory-driven search for an explanation. The scientists in our study, however, sought support for hypotheses in the visually displayed data, and attempted to place the anomaly in the local context of neighboring phenomena. Only hypotheses about expected phenomena were developed at a theoretical level.

There are several possible explanations for this difference. Situational differences in the tasks performed by the participants in these two studies might affect their strategy. For the botanists, categorization was the goal *per se.* Although the astronomers and physicist were performing some categorization tasks, this was done in service of understanding the data as a whole, in order to build a mechanistic theory. The difference in their goals might account for the different strategies they used. Another possibility is that the botanists were getting immediate feedback on their hypotheses, whereas the other scientists had to generate their own feedback. In this sense, the botanists' task is similar to a supervised learning task, whereas the astronomers and physicist were in a situation where learning was unsupervised (Hertz, Krogh, & Palmer, 1991). It is plausible that the uncertainties inherent in this situation can account for the fact that these scientists sought feedback in the empirical data in the display rather than jumping immediately to their theoretical domain knowledge.

Acknowledgments

This research was supported in part by grant 55-7850-00 to the second author from ONR. We thank Wai-Tat Fu and William Liles for comments.

References

Alberdi, E., Sleeman, D. H. & Korpi, M. (2000). Accommodating surprise in taxonomic tasks: The role of expertise. *Cognitive Science, 24*(1), 93-122.

Chinn, C. A., & Brewer, W. F. (1992). Psychological reponses to anomalous data. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Chinn, C. A., & Malhotra, (2001). Epistemologically authentic scientific reasoning. In K. Crowley, C. D. Schunn & T. Okada, (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings.* Mahwah, NJ: Erlbaum.

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & J. Vaid, (Eds.), *Creative thought: An investigation of conceptual structures and processes*. Washington, DC, USA: APA Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). Introduction to the theory of neural computation. Addison-Wesley, Reading, MA.

Kuhn, T. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Lakatos, I. (1976). *Proofs and refutations.* Cambridge, UK: Cambridge University Press.

Kulkarni, D., & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science, 12*(2), 139-175.

Mahoney, M. J., & DeMonbreun, B. G. (1977). Psychology of the scientist: An analysis of problem-solving bias. *Cognitive Therapy and Research, 3,* 229-238.

Mitroff, I. (1974). *The subjective side of science: A philosophical inquiry into the psychology of the Apollo moon scientists.* Amsterdam: Elsevier.

Mynatt, C. R. , Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 29*(1), 85-95.

Nersessian, N. (1999). Model based reasoning in conceptual change. In L. Magnani & N. Nersessian (Eds.), *Model-based reasoning in scientific discovery* (pp. 5-22). New York, NY: Kluwer Academic.

Trickett, S. B., Trafton, J. G., & Schunn, C. D. (2000). Blobs, dipsy-doodles and other funky things: Framework anomalies in exploratory data analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

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