
Theory and Data Interactions of the Scientific Mind: Evidence From the Molecular and the Cognitive Laboratory

Jonathan A. Fugelsang, Courtney B. Stein, Adam E. Green, and Kevin N. Dunbar
Department of Psychological and Brain Sciences
Dartmouth College

Abstract A number of researchers and scholars have stressed the importance of disconfirmation in the quest for the development of scientific knowledge (e.g., Popper, 1959). Paradoxically, studies examining human reasoning in the laboratory have typically found that people display a *confirmation bias* in that they are more likely to seek out and attend to data consistent rather than data inconsistent with their initial theory (Wason, 1968). We examine the strategies that scientists and students use to evaluate data that are either consistent or inconsistent with their expectations. First, we present findings from scientists reasoning “live” in their laboratory meetings. We show that scientists often show an initial reluctance to consider inconsistent data as “real.” However, this initial reluctance is often overcome with repeated observations of the inconsistent data such that they modify their theories to account for the new data. We further examine these issues in a controlled scientific causal thinking simulation specifically developed to examine the reasoning strategies we observed in the natural scientific environment. Like the scientists, we found that participants in our simulation initially displayed a propensity to discount data inconsistent with a theory provided. However, with repeated observations of the inconsistent data, the students, like the scientists, began to see the once anomalous data as “real” and the initial bias to discount that data was significantly diminished.

Science ... warns me to be careful how I adopt a view which jumps with my preconceptions, and to require stronger evidence for such belief than for one to which I was previously hostile. My business is to teach my aspirations to conform themselves to fact, not to try and make facts harmonize with my aspirations. (Huxley, 1860)

These words by Thomas Huxley highlight a phenomenon that scholars have struggled with for centuries – the predisposition of individuals to seek out, interpret, and weight evidence in ways that are consistent with their pre-existing beliefs and expectations, while downplaying or ignoring evidence that is incon-

sistent with their beliefs and expectations. This phenomenon is commonly referred to as the *confirmation bias* (Nickerson, 1998) and is one of the most prevalent sources of inferential error found in human reasoning (Evans, 1989). Researchers from a variety of disciplines, including cognitive psychology (e.g., Bruner, Goodnow, & Austin, 1956; Evans, 1989; Klayman & Ha, 1987; Koriat, Lichtenstein, & Fishhoff, 1980; Mynatt, Doherty, & Tweney, 1977; Wason, 1968), scientific thinking (e.g., Cohen, 1985; Gorman, 1989; Mitroff, 1974; Tweney, 1989; Tweney & Doherty, 1983), judicial reasoning (e.g., Hendry & Shaffer, 1989; Pennington & Hastie, 1993), medical reasoning (e.g., Elstein & Bordage, 1979), and politics (e.g., Healy, 1996) have noted the preponderance of confirmatory-based strategies in human reasoning. One does not have to look far to find examples of confirmation biases depicted in the media. For example, the recent report from the Columbia Accident Investigation Board stated that NASA’s failure to consider relevant data contributed to the recent space shuttle disaster (Sanger, 2003).

A common form of reasoning where confirming and disconfirming strategies are apparent is *causal reasoning*. Indeed, many of the learning and discovery processes in which scientists and nonscientists engage pertain to the development and testing of causal models portraying the relationship between variables of interest (Dunbar, 1995, 2001; Dunbar & Fugelsang, in press). For example, does the Atkins’ diet result in weight loss, smoking cause lung cancer, or aspirin reduce the chance of heart attack? Scientists and nonscientists alike are constantly inundated with claims regarding the causal relationship between such variables.

One’s ability to assess the validity of causal claims is often complicated by the nondeterministic (i.e., probabilistic) nature and the complexity of most cause and effect relationships in the natural environment. For example, the observed relationship between smoking

and lung cancer is probabilistic in nature in that a proportion but not all of those who smoke contract lung cancer, just as a proportion of those who do not smoke contract lung cancer. The multidimensional nature of both causes and effects, the impossibility of accounting for all potential extraneous variables, and errors of data measurement often confound clear delineation of these relationships.

Given the claims that scientists frequently disregard data that are inconsistent with their preferred theories, we wanted to investigate the ways that scientists and nonscientists reason about data that are either consistent or inconsistent with their expectations. Considering the complex environment in which scientists work and the multidimensional nature of the problems often investigated, a complete understanding of the strategies used by scientists can only be gathered by using a multipronged approach including both tightly controlled cognitive experiments and naturalistic observation of scientists reasoning “live” in their laboratories (see Dunbar & Blanchette, 2001 for an example of this research strategy applied to analogical reasoning). In the current research, we apply this two-pronged approach to the study of causal reasoning strategies in scientific inquiry.

Study 1: Scientific Causal Reasoning in the Real World

We first wanted to examine how scientists reasoned about consistent and inconsistent data “live” in their laboratories. We were especially interested in the extent to which scientists applied confirmatory versus disconfirmatory reasoning strategies when receiving data that were either consistent or inconsistent with their predictions. Confirmatory reasoning strategies, in this naturalistic context, can reveal themselves in a number of ways. Do scientists accept consistent data without question? Do scientists spend more time reasoning about consistent or inconsistent data? Are inconsistent data simply explained away as error, or are they closely scrutinized and followed up with further tests of replicability?

Method

Laboratories. Three leading molecular biology laboratories at a prominent U.S. university were analyzed. A detailed account of the three laboratories investigated as well as the precise methodologies used, including explanations of audiotape and videotape transcription and protocol analyses procedures, can be found in Dunbar (1995, 2001). All three laboratories were of similar size and structure, having three to four post-doctoral associates, three to five graduate students, and one or two technicians.

Procedure. Kevin Dunbar interviewed the scientists, attended the laboratory meetings, and read grant proposals and drafts of papers. For the purpose of this investigation, we will focus our analyses on the weekly laboratory meetings, as they constituted a rich source of data representing a wide cross-section of reasoning processes. The laboratory meetings were audiotaped and transcribed. They were then coded along a number of dimensions, including the nature of the experimental finding (e.g., whether it was consistent or inconsistent with expectations), the cognitive operations used (e.g., inductive, deductive, analogical, and causal reasoning), and types of interactions between speakers (e.g., clarification, agreement and elaboration, disagreement, and questioning). Two independent transcribers trained in molecular biology completed all transcriptions.

Results and Discussion

We will focus our analyses on 12 laboratory meetings, four from each of the three molecular biology laboratories. Due to the remarkable similarity between the three laboratories in terms of cognitive operations and experimental outcomes, the data were aggregated for the purpose of analyses. The research programs of these laboratories involved a variety of experimental questions related to how genes control and promote replication in bacteria, parasites, and viruses. For example, one laboratory included in our investigation was conducting various experiments to determine the causal mechanism by which the HIV virus infiltrates the host organism. Consider the following protocol collected where a researcher explains multiple potential mechanisms by which the HIV virus might bind with the host cell.

As you can imagine, for instance, for a cellular mechanism, there is this cellular polymerase that fills in the four bases leaving you with just a small gap... another possibility is that a viral component is what causes this to occur. In this case you can imagine maybe reverse transcriptase is what fills in this four base region. So, for either path, you, now have a finished end provirus.

Here, the researcher clearly outlines two causal hypotheses that formed the basis for a series of experiments. In addition to the theoretically motivated experiments, many of the studies conducted in these laboratories involved the development and testing of new methodologies.

Our main units of analyses are the reasoning strategies that scientists use when faced with data that are consistent or inconsistent with their initial predictions. We will concentrate our analyses on two main aspects

of the data: (1) the frequency of occurrence and the types of causal reasoning strategies elicited by consistent versus inconsistent findings, and (2) the changes in reasoning strategies that occur as a function of replicated inconsistent findings.

Reasoning about consistent and inconsistent data.

The analysis of the 12 laboratory meetings yielded 28 research projects, with 165 experiments, producing a total of 417 results. When the 417 results of the 165 experiments were divided into consistent versus inconsistent findings, we found that over half of the experimental findings were inconsistent with the scientists' predictions (223 out of 417 results). The relatively equal distribution of consistent and inconsistent results permitted a thorough analysis of the different types of causal operations that the scientists undertook as a function of the consistency of their obtained results.

Once a finding was classified, the scientists treated the results in different ways. Consistent results typically led to the next step in a sequence of experiments that was being planned. Inconsistent results, however, prompted a variety of causal reasoning processes. Specifically, scientists developed causal explanations for the inconsistent findings. These causal explanations could be classified into one of two types: (1) methodological or (2) theoretical. The predominant strategy, which occurred for 196 of the 223 inconsistent findings, was to blame the method used in the experiment. In these cases, the scientists would try to find a methodological problem (e.g., wrong incubation temperature) in the experimental methodology. Alternatively, scientists offered theoretical explanations for the data that were inconsistent with their predictions (27 of the 223). Here, the scientists examined either existing theoretical models or revised theoretical models to account for the novel finding.

Changes in reasoning strategies as a function of replicated inconsistent findings. The finding that 12% (27 out of 223) of initial observations of inconsistent findings resulted in theory modification is indicative of a conservative strategy for theory change. Post laboratory meeting interviews suggest that the use of this strategy is based largely on the researchers' knowledge of the high base rate of experimental methodological error. Of the 223 inconsistent findings that occurred, the majority (154 out of 223) were followed up utilizing the same methodology, modified methodologies, or similar control conditions in other experiments. Of those follow-up experiments, 84 resulted in replications of the inconsistent findings. Interestingly, the way that the scientists reasoned about inconsistent findings changed as a function of their repeated occurrence.

When repeated observations of inconsistent findings occurred, scientists began to modify their causal model of how the variables of interest were related. For example, of the 84 anomalous replications, scientists now offered 51 theoretical and only 33 methodological explanations. That is, the plausibility of the once anomalous finding being a legitimate scientific discovery, one that warrants theoretical consideration, was substantially increased.

In summary, the analyses of the causal reasoning strategies in the three molecular biology laboratories have demonstrated that scientists are often reluctant to accept an isolated instance of a finding that is inconsistent with their predictions. However, the inconsistent data are not simply tossed away as error. Rather, in the majority of the cases observed, inconsistent findings were further scrutinized and tested through repeated experimentation. Furthermore, 61% of the replicated inconsistent findings resulted in the scientists re-formulating their original causal theories. Note that this represents a dramatic increase from the 12% of theory modifications that occurred as a function of the initial observation of inconsistent findings.

Study 2: Scientific Causal Reasoning in
the Cognitive Laboratory

We next sought to examine the relationship between one's belief in a causal theory and data consistency in a more controlled setting. To do this, a laboratory equivalent of the scientific environment observed in "real world" molecular biology laboratories was created. In order to devise a nondeterministic environment similar to that observed in the scientists' laboratories, we adapted a methodology commonly used in the cognitive laboratory to measure causal reasoning processes based on probabilistic data. Probabilistic data often take into account the combined role of sufficiency and necessity. The *sufficiency* of a cause is determined by the probability that the effect occurs in the presence of a cause [$P(e/c)$], whereas the *necessity* of a cause is determined by the probability that the effect occurs in the absence of a cause [$P(e/\sim c)$]. The roles of sufficiency and necessity have been featured prominently in contemporary theories of causal thinking (e.g., Cheng, 1997; Novick & Cheng, in press; White, 2002) and numerous experiments conducted in the psychological laboratory have supported the assumption that people do indeed form causal models based to a large degree on the observed contingency (i.e., covariation or probabilistic relationship) between variables (e.g., Cheng & Novick, 1990; Lober & Shanks, 2000; Spellman, 1996; White, 2002).

Do students, like the scientists, initially show a reluctance to accept a strong experimental finding as

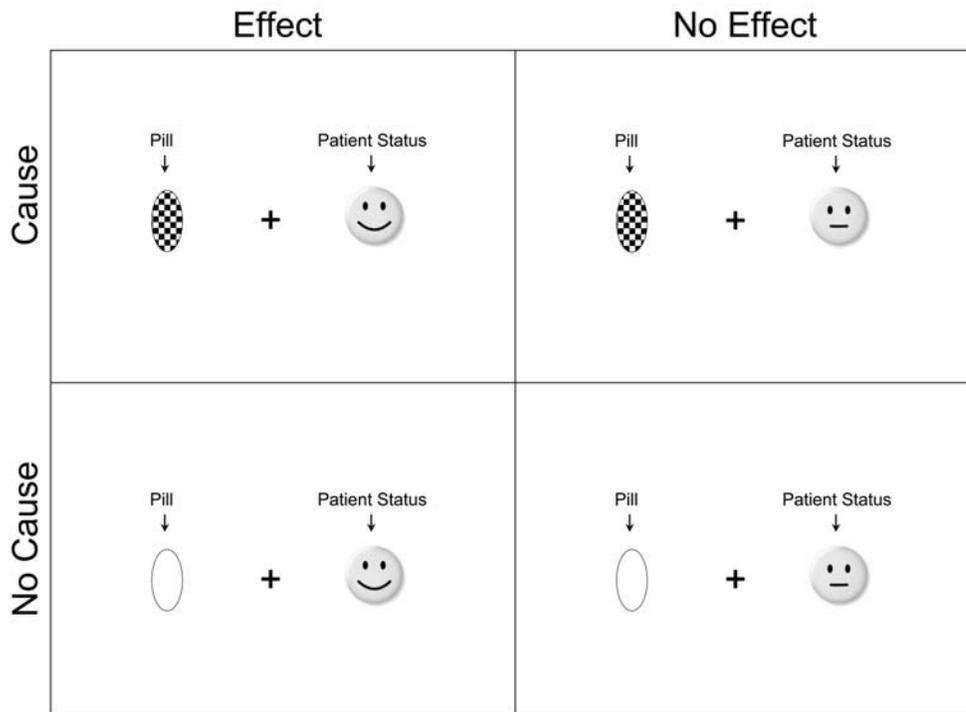


Figure 1. Example stimuli representing the four possible combinations of the candidate cause (checked pill vs. white pill) and effect (happiness vs. neutral emotion). Note that the stimuli in the actual experiments utilized a red pill and a blue pill in the place of the checked pill and the white pill.

causally valid if there is no plausible causal theory supporting it? Do repeated observations of inconsistent data influence students' willingness to accept inconsistent data as causally relevant? The following experiment addresses these issues by manipulating the plausibility of a causal theory, the degree to which the data are consistent or inconsistent with the theory, and the amount of data available.

Methods

Participants. Thirty-two participants (25 females and 7 males, mean age = 21.31 years) took part in the study and were paid US\$10. Informed written consent for all participants was obtained prior to the experiment in accordance with the guidelines established by the Committee for the Protection of Human Subjects at Dartmouth College.

Design and apparatus. This experiment was a 3 x 2 x 4 within subjects design with plausibility of the causal theory (no direct causal link predicted, neutral, and direct causal link predicted), strength of covariation-based data (weak and strong covariation), and sample size (10, 20, 30, and 40 data trials) as within-subject variables. All stimuli were presented on a G3

iMac computer running PsyScope 2.5.1 software (Cohen, MacWhinney, Flatt, & Provost, 1993).

Materials and procedure. The plausibility of the theory of action of a drug and whether the data were consistent or inconsistent with the theory were varied. The plausibility of a theory was manipulated by presenting participants with a brief introductory statement that contained either (1) a direct plausible causal mechanism of action linking a red pill to a mood outcome, (2) *no* direct causal mechanism of action linking a red pill to a mood outcome (i.e., analogous to an experimental *control* condition in the "real world" environment), or (3) a neutral causal mechanism of action (see Appendix). This level of the plausibility variable will be referred to as *plausible*, *implausible*, and *neutral* theories, respectively. The causal mechanisms consisted of biological agents in order to create a situation that was roughly analogous to those observed in the three molecular biology laboratories.

Data were then provided to participants in a trial-by-trial format where they viewed 40 trials of data for each causal theory provided. These data were presented in combinations of the cause (a *red pill* or a *blue pill*) and the effect (*happiness* or *neutral* outcome) co-

occurring. Figure 1 presents a graphical depiction of these four event types. Under some conditions, the *red pill* and *happiness* covaried strongly, under other conditions, the *red pill* and *happiness* covaried weakly. This was accomplished by varying the frequency with which each of the four event types (*red pill/happiness*, *red pill/neutral*, *blue pill/happiness*, *blue pill/neutral*) occurred. In the strong covariation condition, DP_c [i.e., $P(e/c) - P(e/\sim c)$] was equal to .7; for the weak covariation condition, DP_c was equal to .3. The marginal totals (i.e., total number of observations where the cause was present or absent) were set at 40 for both levels of DP_c . Note that strong covariation-based data following a plausible causal theory and weak covariation-based data following an implausible causal theory both constitute *consistent* data, whereas weak covariation-based data following a plausible theory and strong covariation-based data following an implausible theory both constitute *inconsistent* data.

Participants advanced each trial of data by pressing the space bar on the computer keyboard. Four times throughout each data-testing period, participants were asked to make a rating about how probable they think it is that the *red pill* caused the happiness using a scale that ranged from 1 (Low) to 5 (High). These ratings were made after 10, 20, 30, and 40 trials of patient data. Participants were instructed to treat the 40 trials of data as cumulative. This procedure was repeated six times: once for each level of the theory plausibility and covariation-based data manipulations. Note that subjects were not given any information about the blue pill and were not asked to make any ratings about the blue pill. The order in which each causal theory was presented, and the order in which each event type within each testing period occurred was random.

Results and Discussion

The results will be presented in two sections. The first section presents the omnibus analyses of theory plausibility (implausible, neutral, and plausible), strength of the covariation-based data (strong and weak), and sample size (10, 20, 30, and 40 patient trials). The second section presents the effect of sample size on the interplay between theory plausibility and strength of the covariation-based data. The alpha level for all statistical tests was set at .05 (two-tailed) unless otherwise stated. Effect size estimates were computed using partial η^2 .

Theory, data, and sample size. Figure 2 presents the mean causal ratings for the three theory plausibility levels, the two covariation levels, and the four sample sizes. The causal ratings were analyzed using a 3 x 2 x 4 (Plausibility Level x Data Strength x Sample Size)

repeated measures ANOVA. As expected, there was a main effect of theory plausibility $F(2,62) = 21.87$, $MSE = 1.93$, $\eta^2 = .41$, where causal ratings were higher for conditions containing a plausible causal mechanism ($M = 3.31$) than either a neutral ($M = 3.07$) or an implausible causal mechanism ($M = 2.52$). Individual paired t-tests revealed that all three means were reliably different from each other (smallest $t = 2.56$). In addition, causal ratings were also higher when the covariation-based data were strong ($M = 3.42$) than when the covariation-based data were weak ($M = 2.10$), $F(1,31) = 269.11$, $MSE = 1.96$, $\eta^2 = .90$. Importantly, there was also a Plausibility x Data Strength interaction, $F(2,62) = 9.83$, $MSE = .94$, $\eta^2 = .24$, where the effect of the data strength manipulation increased parametrically as a function of the plausibility of the causal theory. Specifically, the effect of data strength was largest when the theory being tested was plausible (M difference = 2.03) as compared to when the theory was neutral (M difference = 1.67), or implausible (M difference = 1.27). These three interaction terms were all reliably different from each other (smallest $t = 2.22$). There was also a main effect of sample size, $F(3,93) = 3.46$, $MSE = 0.69$, $\eta^2 = .10$, where causal ratings increased as a function of increasing sample size. Individual paired t-tests revealed that the locus of the main effect was the increase in magnitude of the causal ratings between a sample size of 10 ($M = 2.88$) and 40 ($M = 3.13$), $t(31) = 2.91$, $SE = .09$; all other comparisons were not reliably different (largest $t = 1.91$).

Effect of sample size on the interplay between theory and data. We next wanted to examine the degree to which the interaction between plausibility and data strength varied as a function of sample size. To simplify the description of these analyses, the four sets of analyses (one for each sample size) will be summarized together in terms of (1) main effects of theory plausibility, (2) main effects of data strength, and (3) the presence or absence of a Plausibility x Data Strength interaction. First, main effects of theory plausibility (smallest $F = 9.92$, $\eta^2 = .24$), and data strength (smallest $F = 102.08$, $\eta^2 = .78$), were found for all sample sizes. Importantly, however, the Plausibility x Data Strength interaction was significant in the 10, 20, and 30 trial sample sizes (smallest $F = 4.89$, $\eta^2 = .14$) but not in the 40-trial sample size, $F(2,62) = 1.59$, $MSE = .44$, $\eta^2 = .05$. Analyses of participants' covariation discrimination (i.e., difference in causal ratings for high minus low covariation) for each theory as a function of sample size revealed that participants' covariation discrimination increased as a function of the sample size manipulation (i.e., 10, 20, 30, and 40 patient trials) for low plausible theories, $F(3,93) = 3.16$, $MSE = 1.05$, $\eta^2 =$

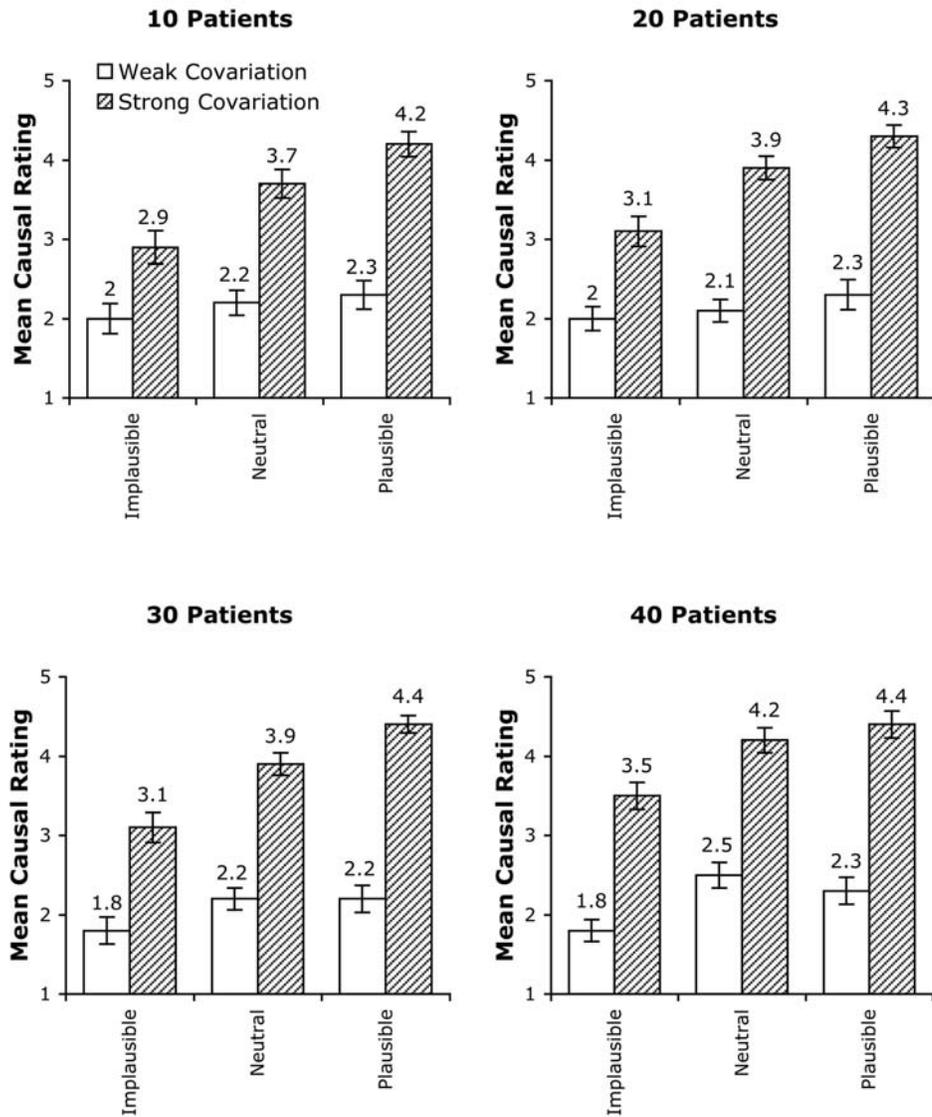


Figure 2. Mean causal ratings for the three theory plausibility levels (implausible, neutral, and plausible), the two covariation conditions (weak and strong), and the four sample sizes (10, 20, 30, and 40 patients).

.09) whereas participants' covariation discrimination for the neutral and highly plausible theories remained unchanged as a function of sample size (both $F_s < 1$).

In summary, the dependence of the Theory Plausibility x Data Strength interaction on sample size can be interpreted as evidence for a multifaceted account of scientific causal thinking. Specifically, when there was a strong relationship observed in the context where none was expected (i.e., implausible theory), the data appeared to have been initially discounted by the participants. In contrast, when a strong relationship was observed in the context of a theory that predicted a relationship to exist (i.e., plausible theory), the data

were given more weight. This propensity to discount data inconsistent with an implausible theory was modulated by the amount of data present. Here, many replications (i.e., 40-patient trial condition) of strong data for an implausible causal theory (i.e., inconsistent data) increased participants' ratings of causality. These findings are consistent with the data observed in the real-world laboratories of the molecular biologists. There, too, repeated observations of inconsistent data resulted in modifications of original theories and thus increased acceptance of the inconsistent data as "real" and nonanomalous.

General Discussion

In the two studies reported, we have shown that scientists and nonscientists display similar strategies when dealing with data that are consistent or inconsistent with their causal theories. While data consistent with a theory are met with little scrutiny, data inconsistent with a theory are initially met with skepticism, resulting in primarily methodological explanations by scientists in the molecular biology laboratories and low causal ratings from students participating in the scientific reasoning simulation. However, this initial tendency to accept data consistent with a theory and discount data inconsistent with a theory can be overcome by replications of the inconsistent data.

The initial inclination to question data inconsistent with a theory does not necessarily represent a faulty reasoning strategy in a practical sense. Due to experimenter error and methodological inconsistencies from lab to lab, anomalous findings may, and often do, occur for a number of theoretically insignificant reasons. Initial skepticism of inconsistent findings can act as a failsafe against prematurely modifying one's theoretical understanding of the variables under study. Indeed, Baker and Dunbar (2000) have shown that scientists often include both "known standard" and "baseline" control conditions for this very reason.

Relevance to Models of Scientific Causal Thinking

These data are broadly consistent with findings observed in the inductive reasoning (e.g., Gorman & Gorman, 1984; Wason, 1968), hypothesis testing (e.g., Klayman & Ha, 1987), and the scientific thinking literature (Tweney, 1989; Tweney & Doherty, 1983). Here, it has been shown that people initially adopt a confirmatory reasoning-based strategy but turn to disconfirmatory strategies when confirmatory-based strategies fail. These data also provide a theoretical extension to recent models of causal thinking that incorporate theory and data interactions (Fugelsang & Thompson, 2000, 2002, 2003). Specifically, we provide an account of how data replicability may influence the interplay between theory and data. Here, we show that data inconsistent with an implausible theory are initially met with skepticism. However, through the course of repeated observations of the inconsistent data, people begin to modify their initial theory and, as such, increase their causal ratings. This initial reluctance and subsequent re-theorizing can be thought of as a useful heuristic in that it serves two primary decision-making purposes. First, it prevents people from prematurely accepting findings that may be spurious. Indeed, if one modified his/her theoretical beliefs for every occurrence of data that are inconsistent with a theory, one would continually need to modify his/her knowledge

and be unable to form any strong causal impressions. Secondly, it permits the updating of theories and the development of knowledge through repeated observation. The finding that one's knowledge can be modified with extensive replications provides an optimistic view of causal reasoning heuristics used by scientists and nonscientists alike.

One especially interesting finding, one that warrants further investigation, is the asymmetry observed between the effects of theory plausibility and sample size when data were inconsistent with a prescribed theory. Specifically, it appears that participants in our study were prepared to give a lot of weight to null findings with few trials when they had a theory that led them to expect a causal relationship. In contrast, participants were more reluctant (i.e., required more trials) to accept positive findings when they were led to expect no causal relationship. One possible explanation for this asymmetry is that people's beliefs in the capacity of a potential cause may be independently influenced by their strength or personal conviction in those beliefs. Indeed, Poletiek and Berndsen (2000) and Koehler (1993) have demonstrated that the subjective *value* and *strength* of personal beliefs may alter the strategies that participants employ when testing hypotheses and judging the quality of data. Here, the extent to which participants are willing to renounce prior beliefs in a causal theory may be related to the strength of those beliefs independent of their plausibility.

In summary, we have provided evidence that scientists reasoning "live" in their laboratories and students in a scientific reasoning simulation both demonstrate an initial reluctance to consider data inconsistent with their predictions. On the surface, these findings are consistent with traditional accounts of confirmation bias that argue that people possess an inherent disposition to downplay data inconsistent with their expectations. We have provided an extension to this account that incorporates the amount of inconsistent data present. Here, we show that the confirmation bias is significantly reduced under situations where people receive a preponderance of inconsistent data. As these processes are surely influenced by motivational factors, level of expertise of the reasoner, and knowledge domain, future research should examine the extent to which this heuristic is modulated by both individual difference variables among the reasoners, and situational factors in the reasoning environment.

Research reported in this paper has been funded by grants from Dartmouth College, The Spencer Foundation, Natural Sciences and Engineering Research Council of Canada, McGill University, and The Office of Naval

Research. Address correspondence to Jonathan A. Fugelsang or Kevin N. Dunbar, Psychological and Brain Sciences, 6207 Moore Hall, Dartmouth College, Hanover, New Hampshire, 03755 (E-mail: jonathan.a.fugelsang@dartmouth.edu; kevin.n.dunbar@dartmouth.edu).

References

- Baker, L. M., & Dunbar, K. (2000). Experimental design heuristics for scientific discovery: The use of "baseline" and "known standard" controls. *International Journal of Human Computer Studies*, *53*, 335-349.
- Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.
- Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*, 545-567.
- Cohen, I. B. (1985). *Revolution in science*. Cambridge, MA: Harvard University Press.
- Cohen J. D., MacWhinney B., Flatt M., & Provost J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers*, *25*, 257-271.
- Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. Davidson (Eds.), *Mechanisms of insight* (pp. 365-395). Cambridge, MA: MIT press.
- Dunbar, K. (2001). The analogical paradox: Why analogy is so easy in naturalistic settings, yet so difficult in the psychology laboratory. In D. Gentner, K. J. Holyoak, & B. Kokinov, *Analogy: Perspectives from cognitive science* (pp. 313-334). Cambridge, MA: MIT press.
- Dunbar, K., & Blanchette, I. (2001). The *InVivo/InVitro* approach to cognition: The case of analogy. *Trends in Cognitive Sciences*, *5*, 334-339.
- Dunbar, K., & Fugelsang, J. (in press). Causal thinking in science: How scientists and students interpret the unexpected. To appear in M. E. Gorman, A. Kincannon, D. Gooding, & R. D. Tweney (Eds.), *New directions in scientific and technical thinking*. Mahwah, NJ: Lawrence Erlbaum.
- Elstein, A. S., & Bordage, G. (1979). Psychology of clinical reasoning. In G. Stone, F. Cohen, & N. Adler (Eds.), *Health psychology: A handbook* (pp. 333-367). San Francisco, CA: Jossey-Bass.
- Evans, J. St. B. T. (1989). *Bias in human reasoning*. Hillsdale, NJ: Erlbaum.
- Fugelsang, J., & Thompson, V. (2000). Strategy selection in causal reasoning: When beliefs and covariation collide. *Canadian Journal of Experimental Psychology*, *54*, 13-32.
- Fugelsang, J., & Thompson, V. (2002). Foundations of human causal reasoning. In S. Shohov (Ed.), *Advances in psychology research: Vol. 12* (pp. 83-101). Huntington, NY: Nova Science Publishers, Inc.
- Fugelsang, J., & Thompson, V. (2003). A dual-process model of belief and evidence interactions in causal reasoning. *Memory & Cognition*, *31*, 800-815.
- Gorman, M. E. (1989). Error, falsification and scientific inference: An experimental investigation. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *41A*, 385-412.
- 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: Human Experimental Psychology*, *36A*, 629-648.
- Healy, B. (1996, July 3). The dangers of trial by Dingell. *The New York Times*.
- Hendry, S. H., & Shaffer, D. R. (1989). On testifying in one's own behalf: Interactive effects of evidential strength and defendant's testimonial demeanor on jurors' decisions. *Journal of Applied Psychology*, *74*, 539-545.
- Huxley, T. H. (1860). The origin of species, in *Collected Essays, Vol. 2* (pp. 71-79). London: Macmillan.
- Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211-228.
- Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, *56*, 28-55.
- Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning & Memory*, *6*, 107-118.
- Lober, K., & Shanks, D. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review*, *107*, 195-212.
- Mitroff, I. (1974). *The subjective side of science*. Amsterdam, NL: Elsevier.
- Mynatt, B. T., 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*, 85-95.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomena in many guises. *Review of General Psychology*, *2*, 175-220.
- Novick, L., & Cheng, P. (2004). Assessing interactive causal inference. *Psychological Review*, *111*, 455-485.
- Pennington, N., & Hastie, R. (1993). The story model of juror decision making. In R. Hastie (Ed.), *Inside the juror: The psychology of jury decision making* (pp. 192-221). New York: Cambridge University Press.
- Poletiek, F. H., & Berndsen, M. (2000). Hypothesis testing as risk behaviour with regard to beliefs. *Journal of Behavioural Decision Making*, *13*, 107-123.
- Popper, K. R. (1959). *The logic of scientific discovery*. London, UK: Hutchinson.

- Sanger, D. E. (2003, August 27). Inertia and indecision in NASA. *The New York Times*.
- Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative causes. *Psychological Science*, 7, 337-342.
- Tweney, R. D. (1989). A framework for the cognitive psychology of science. In B. Gholson, A. Houts, R. A. Neimeyer, & W. Shadish (Eds.), *Psychology of science and metascience* (pp. 342-366). Cambridge, MA: Cambridge University Press.
- Tweney, R. D., & Doherty, M. E. (1983). Rationality and the psychology of inference. *Synthese*, 57, 139-161.
- Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20, 273-281.
- White, P. A. (2002). Perceiving a strong causal relation in a weak contingency: Further investigation of the evidential evaluation model of causal judgment. *The Quarterly Journal of Experimental Psychology*, 55A, 97-114.

Appendix

Stimuli used in scientific causal reasoning simulation. The stimuli were pretested for plausibility with an independent sample of 23 subjects. This separate group of participants was simply asked to rate the degree to which the presented pill could cause the elevation of mood in a sample of patients. They were not provided with any covariation information. Following each scenario is the mean pretested plausibility rating (*Plausibility*) and standard deviation (*SD*).

Plausible Mechanism

- (1) Past research has demonstrated that people's feelings of happiness are directly related to the level of *serotonin* in the brain. The red pill is a "selective serotonin reuptake inhibitor." This pill blocks the recycling process for the *serotonin*, which then keeps more of this neurotransmitter in the brain available to communicate with other nerve cells. (*Plausibility* = 8.18, *SD* = 1.07).
- (2) Past research has demonstrated that people's feelings of happiness are directly related to the level of *norepinephrine* in the brain. The red pill is a "monoamine oxidase inhibitor." Monoamine oxidase is an enzyme that breaks down *norepinephrine* in the brain. Monoamine oxidase inhibitors inhibit this enzyme, thus allowing a greater supply of this neurotransmitter to remain available in the brain. (*Plausibility* = 8.22, *SD* = .95).

Implausible Mechanism

- (1) Past research has demonstrated that the growth of small amounts of the bacteria *staphylococcus* in the body has no direct link to people's feelings of happiness. The red pill is a "topoisomerase inhibitor." Topoisomerase is an enzyme that is necessary for the reproduction of *staphylococcus* in the body. "Topoisomerase inhibitors" inhibit this enzyme, thus restricting the ability of *staphylococcus* to replicate. (*Plausibility* = 1.17, *SD* = 1.54).
- (2) Past research has demonstrated that the growth of small amounts of the bacteria *clostridium* in the body has no direct link to people's feelings of happiness. The red pill is a "protein binder." The cell walls of bacteria are continuously expanding through the synthesis of proteins and amino acids. In order for a bacteria cell to flourish and reproduce, the cell wall must be able to expand with the growing interior. "Protein binders" bind to specific amino acids and proteins, thus inhibiting the cell wall of *clostridium* to synthesize. (*Plausibility* = .78, *SD* = .80).

Neutral Mechanism

The active chemical agents of the red pill are unknown.

Sommaire

Les auteurs de la présente enquête se penchent sur les stratégies utilisées par scientifiques et étudiants pour évaluer des données qui sont soit conformes soit non conformes à une théorie causale. L'examen est fondé tant sur l'observation naturalistique que sur des méthodes d'expérimentation contrôlée. Dans un premier temps, les auteurs présentent les constatations de scientifiques qui raisonnent « de vive voix » à des réunions du personnel de laboratoires. Ils découvrent ainsi des stratégies utilisées par les scientifiques pour déterminer la validité de données qui sont soit conformes soit non conformes à leurs théories. L'analyse repose principalement sur les stratégies de raisonnement auxquelles les scientifiques ont recours face à des données conformes ou non conformes à leurs prédictions initiales. Les auteurs s'intéressent à deux principaux aspects des données : (1) la fréquence de l'occurrence et les types de stratégies de raisonnement causal suscitées par des constatations conformes par opposition à des constatations non conformes; (2) les changements aux stratégies de raisonnement découlant de la répétition d'observations non conformes. L'analyse des données précitées révèle que les scientifiques sont souvent réticents à accepter une constatation isolée non conforme à leurs prédictions théoriques. Par contre, ils n'écartent pas simplement les données non conformes, les jugeant erronées. Dans la majorité des cas

observés, des constatations non conformes ont été examinées et testées plus à fond par des expériences répétées. Les auteurs ont approfondi l'examen de la question au moyen d'une simulation contrôlée de réflexion scientifique expressément conçue pour analyser les stratégies de raisonnement qu'ils avaient observées en milieu scientifique naturel. À l'instar des scientifiques, ils ont constaté que les participants à la simulation ont manifesté, dans les premiers temps, une tendance à écarter les données non conformes à leurs théories. Toutefois, à force d'observations répétées de données non conformes, les étudiants, à la manière des scientifiques, ont commencé à percevoir comme « réelles » les données autrefois jugées des anomalies, et le biais initial qui entravait le rejet des données en était diminué de façon appréciable. À première vue, ces constatations se situent dans la ligne des explications classiques du biais en faveur de la confirmation, selon lesquelles les gens sont foncièrement disposés à minimiser l'importance des données non conformes à leurs attentes et à se concentrer plutôt sur celles qui les avèrent. Les auteurs ont poussé plus loin ces explications et tenu compte du volume de données non conformes en cause. Ils ont montré, dans les circonstances, que le biais en faveur de la confirmation est sensiblement réduit dans des situations où les gens reçoivent des données dont la plupart sont non conformes.