When half right is not half bad: Hypothesis testing under conditions of uncertainty and complexity

Joshua Klayman

Center for Decision Research
Graduate School of Business, University of Chicago

Analyses of scientific reasoning, from computer simulation (e.g., Langley et al., 1987) to biographical analyses of famous scientists (e.g., Tweney, 1985), often rely on a prototypical model of the hard sciences, especially physics. On the basis of this prototype, scientific inquiry has usually been modeled in terms of the discovery of laws, like Newton's, Boyle's or Ohm's--the kinds of simple formulae learned in introductory science classes. Although simple, each law explains a broad class of events or relations. These laws have exceptions and complications, especially in exotic conditions, but basically, F = ma for all kinds of masses and all kinds of forces, and PV = nRT for all kinds of gasses in all kinds of containers. The laws are, in principle, deterministically correct, and, within the bounds of measurement error, the data about them are consistent and unambiguous.

Unfortunately, this model of science is not a good representation of hypothesis testing and scientific reasoning in informal settings. In general, the subjects of everyday reasoning (e.g., the behavior of children, the performance of automobiles, the judgments of editors) are not subject to simple explanations. In a lawful science, even slight discrepancies from the law are matters to be reckoned with. With intuitive theories, half right is not half bad: One is often pleased with an explanatory value noticeably better than zero, and the domain of applicability of such theories is usually quite restricted (e.g., one child or one automobile). Intuitive scientists must also deal with considerable ambiguity in the relations between hypotheses and data. Typically, the magnitude of the measurement error is nearly the same as the magnitude of the effects under study. Hypotheses have ambiguous implications about the phenomena that should or should not be observed, and observed data have ambiguous implications as to how or whether a hypothesis should be revised.

The fundamental difference between the law-based prototype of science and the task of reasoning in daily life is the degree of uncertainty and complexity in the environment. By uncertainty, I mean that phenomena must be regarded as probabilistic, and not subject to complete explanation or prediction. By complexity, I mean that phenomena arise from the simultaneous influence of numerous and often inscrutable contributory factors. Although I focus on informal reasoning, many professional scientists also face high uncertainty and complexity, particularly in "soft" or "inexact" fields, like the social sciences, or in newer, less well-established domains such as high-temperature superconductors today.

Uncertainty and complexity have important implications for the kinds of hypotheses people form. In lawful domains, it may be a reasonable approximation to say that scientists search for the "true" explanation. In inexact domains, no hypothesis is expected to yield nearly perfect prediction or nearly complete explanation. Instead, the hypothesis may state that there is an association between two things, or that a certain factor should have a significant effect on a behavior of interest. The goal is not to discover the right rule or law. The goal may instead be the development of a theory that is "pretty good" according to domain-specific standards, or one that is significantly better than the previous hypothesis, or the goal may be merely to achieve predictive accuracy better than chance.

Uncertainty and complexity have parallel effects on the process of testing hypotheses. With probabilistic hypotheses, there

are no logically determined "critical" tests, nor any logical mandate as to how to modify the hypothesis in response to new data. Thus, the processes of hypothesis testing and revision become matters of accumulation and interpretation of evidence, and a process of zeroing in on a hypothesis that meets the hypothesis-tester's goal.

There is a fairly extensive body of research on the psychological processes of hypothesis testing, and this work has produced a number of interesting findings concerning the abilities and failings of human hypothesis testers (see Klayman & Ha, 1987; Nisbett & Ross, 1980; Wason & Johnson-Laird, 1972). In this paper, however, I focus on an area of research not usually associated with the study of scientific reasoning. This research has gone under a variety of names, but can be referred to generically as *cue learning*. Cue-learning tasks require the subject to make judgments based on one or more cues that provide only partial and imperfect information, and to use feedback to try to improve the accuracy of those judgments. Thus, cue learning captures the flavor of everyday reasoning better than many hypothesis testing tasks.

Cue learning

The development of cue learning in the 1950's was based on Egon Brunswik's (1956) principle of "probabilistic functionalism," the principle that judgments in natural environments must be derived from a combination of multiple, imperfect cues. Thus, the central goal of the paradigm is to study how people learn to relate cues to judgments in probabilistic environments. Cue-learning tasks have three basic elements: a criterion (something the subject must learn to predict or estimate), cues (information from which to make the estimate) and feedback (information about the accuracy of the estimates made). As in natural learning environments, the criterion is not fully predictable from the available cues. For example, the criterion value, Y, might be determined by the formula $Y = 1/3\ A + 2/3\ B + C + e$, where A, B, and C are cues, and e is a number from a random number generator.

Cue learning actually encompasses several different kinds of learning. First, there is the matter of "cue discovery" (Klayman, 1988), the process of finding valid cues to use in making predictions. Then, given a set of cues, there is the matter of how those cues should be combined. This includes determining whether effects are additive or multiplicative, and the relative importance weight to give each cue. Then, there is the task of determining the form of the different cue-criterion functions. For example, it may be that, other things being equal, the criterion has a positive linear relation to cue A, a negative linear relation to cue B, and a U-shaped relation to C.

Cue learning captures some of the complexity and uncertainty of real-life hypothesis testing. The behavior of interest is a function of a number of simultaneous factors, there is only a corelational association between any cue and the criterion, and all the available cues, taken together, cannot completely predict or explain the dependent measure. Furthermore, the feedback one gets is ambiguous in the sense that a discrepant finding may reveal an inaccuracy in the model you are using, or it may be attributable to random error; if a change is indicated, it is not clear just what the change should be. Thus, it should perhaps be no surprise that many cue-learning studies have found that it is difficult for people to learn from feedback in such situations (see Brehmer, 1980; Klayman, in press).

Hypotheses in cue learning

During the first couple of decades of cue-learning research, not much attention was paid to the matter of hypotheses. Cue learning was regarded as a process of learning to associate certain criterion values with certain values of each cue,

along with some process of averaging and interpolation. However, there is a growing body of evidence about the crucial role that hypotheses play in learning in complex and uncertain environments. In this regard, there is now an important bridge between cue-learning research and more mainstream research in learning and scientific reasoning.

Where do the hypotheses come from? Except when the learning task is presented abstractly (e.g., with cues identified only as *A*, *B*, and *C*), hypotheses will of course be derived from the learner's knowledge and theories about the causal structure of the environment. World knowledge may suggest specific functions (e.g., that the relation between effort and performance in a task is a positive one, with diminishing returns), or provide more general hints (e.g., to look at personality variables in this situation). At the most general level, one may fall back on general cues to causality such as temporal and spatial proximity (see Einhorn & Hogarth, 1986).

On the other hand, there is also evidence of a general default hierarchy of hypotheses that follows a sort of principle of intuitive parsimony. Given several cues from which to make judgments, subjects hypothesize mostly about how the cues ought to be combined, and they seem to pay little attention to the matter of cue-criterion functions (Brehmer, 1987). However, if they use the cues to make estimates in the meantime, they *must* make some assumptions about the underlying functions, at least de facto. Brehmer found that subjects' responses implied a default assumption of simple linear cue-criterion relations. This is also the most common initial hypothesis about the cue-criterion function in one-cue tasks (Brehmer, 1974). In a task involving cue discovery, in which the set of useful cues was not fully specified in advance (Klayman, 1988), subjects' hypotheses were almost exclusively in the form of "the more of this, the more of [or the less of] that." Subjects seldom expressed any hypotheses about how *much* of this or how much of that, and there were few hypotheses about interactions among cues. Responses implied a default assumption of linear cues that combined additively.

There is little evidence that intuitive parsimony is a conscious strategic principle. Rather, it can be viewed as an outcome of the feedback-encoding process. A simple way to encode feedback is to observe that a change in a cue tends to be associated with some direction of change in the criterion ("this one had more achievement motivation and did worse on the test"). This level of encoding permits cue discovery, since the learner could perceive the existence of an effect. Slightly more complex encoding might include some information about the magnitude of change ("...a lot more achievement motivation and did a little worse..."). This yields information about the average magnitude of the relationship, but nothing about its shape. The only way to recognize a nonlinear relation is to keep track of the relation between the *magnitude of changes* and the *absolute magnitude* of the cue. For example, the relation $Y = \log(X)$ implies that the larger X is, the smaller the change in Y with a given change in X. Similarly, perception of interactions would require encoding the relation between X_1 and Y as a function of X_2. Nonmonotonic functions and disordinal interactions may be particularly hard to learn, because even the *direction* of change will be observed to vary, and the learner may conclude that no consistent relationship exists.

Evidence from a number of cue-learning studies supports this ordering of task difficulty: People learn the identity of cues before they learn how best to combine them; they learn additive combinations more easily than others; and they learn linear relations more easily than nonlinear ones (see Klayman, in press). Part of the difficulty may be that people simply fail to consider hypotheses further down their hierarchy, but it is also the case that the more complex functions are simply more

difficult to perceive (Brehmer, 1980). The default hierarchy of hypotheses also interacts with world knowledge. On the one hand, some of the more complex functions, such as nonmonotonic cues and disordinal interactions, may be learnable if one has a prior hypothesis to guide the encoding of the data. On the other hand, people seem prone to encode their world knowledge in terms of simpler functions and combinations as well (Klayman, in press; Sniezek, 1986).

Testing hypotheses

In the law-discovery model of reasoning, investigators can reasonably expect to determine whether their hypotheses are right or wrong. (At least, following Popper (1959), you should be able to determine whether or not your hypothesis has been falsified yet). In cue learning, though, the goal is not to find out if the hypothesis is wrong (because it always is), but where it is wrong, and how it might be fixed. Hypothesis testing and revision is thus a process of gradual refinement, starting with general ideas about the types of things to consider, and moving to more complete and specific (and hopefully more accurate) hypotheses (see, e.g., Klayman, 1988; Klahr & Dunbar, 1988; Lakatos, 1978).

How then is feedback used to test hypotheses derived from world knowledge, previous feedback, and default rules? Research on hypothesis-testing behavior suggests that hypothesis testing under conditions of uncertainty and complexity is a very difficult task. A number of these difficulties have come under the rubric of "perseverance of beliefs" (Ross & Lepper, 1980) or "confirmation bias" (Fischhoff & Beyth-Marom, 1983; Klayman & Ha, 1987). For example, people tend to interpret ambiguous evidence in a way that favors their current hypothesis. In a probabilistic environment, feedback is always ambiguous is that it is never clear whether deviations from the expected are meaningful or "just" random. People may "immunize" their hypotheses, by accepting results that conform to their hypotheses, while attributing unexpected findings to random error more than is justified (Gorman, 1986). People also use a "positive test strategy" in which attention is focused on the ability of the current hypothesis to predict and explain observed events, with little consideration given to possible alternative hypotheses (Klayman & Ha, 1987, 1988).

The general implication of these aspects of hypothesis testing is that subjects will be slow to reject early hypotheses. This need not always be the case, however. In some situations, people seem very prone to changing hypotheses, and may hurt themselves by rejecting good ones. This will happen if learners have unrealistic expectations about how good a good hypothesis ought to be, i.e., if they underestimate the impact of hidden variables and random error. A number of studies have documented people's tendency to underestimate the role of chance (see Langer, 1975, for example), especially in the absence of world knowledge about the underlying mechanisms (Nisbett et al., 1983). The result can be a fickle hypothesis tester, who rejects and replaces hypotheses on the basis of insufficient negative evidence. This pattern has been observed in a variety of learning studies (Brehmer, 1980; Mynatt, Doherty & Tweney, 1978).

Hypotheses and learning from feedback

The real object of cue learning, and learning from experience in general, is not so much to test hypotheses, but to revise and improve them. From the above discussion, it might appear that when data meet hypothesis, the prospects for appropriate learning are poor. However, the use of feedback to choose and revise hypotheses can have different outcomes, depending on the relation between hypotheses and incoming data.

Not surprisingly, people make more accurate judgments when their hypotheses are

congruent with the data, for example, when the cue labeled "monthly debt" is negatively related to "credit rating" (Muchinsky and Dudycha, 1975). One straightforward explanation for this finding is that people do not *need* to learn from outcome feedback if they already have appropriate hypotheses. However, evidence also suggests that congruent hypotheses can facilitate subsequent learning. For example, Camerer (1981) found that subjects learned to use a multiplicative interaction present in outcome feedback when dimensions were labeled in a way that suggested the interaction. (MBA students perceived an interaction between price changes and trade volume in predicting a commodities market.) In contrast, subjects who were not given feedback did not manifest any such interaction in their estimates, nor did subjects given feedback with unlabeled cues.

Without a concrete hypothesis, subjects face the task of learning the associations between myriad cue values and a whole range of criterion values. Uncertainty and complexity make this abstraction process all the more difficult. An appropriate hypothesis can provide a useful way to organize feedback in encoding, aggregation, and recall. Data can be encoded as supporting or contradicting the hypothesis, and past experiences can be summarized in terms of a limited number of prior hypotheses, rather than a large number of individual feedback data. If a basic hypothesis about a cue seems valid, learners may then be able to use feedback to refine their ideas about the shape of the function and its relation to other cues.

But what if feedback in a learning situation *contradicts* the expectations you bring to it? One might expect misleading hypotheses to seriously interfere with learning, since good but counterintuitive hypotheses may never be considered, and poor but plausible hypotheses may persevere. Indeed, several studies have found that tasks that elicit inaccurate hypotheses are as hard to learn as purely abstract ones (Miller, 1971; Camerer, 1981) or harder (Adelman, 1981). On the other hand, some studies find that the information-processing benefits of a concrete hypothesis may even override misleading expectations. For example, Muchinsky and Dudycha (1975) and Sniezek (1986) report that subjects learned meaningfully labelled relations better than abstract ones even when the data seemed anomalous. In such cases, subjects ad libbed new hypotheses or reinterpreted the data, and then used the new interpretations to encode subsequent feedback. For example, some of Sniezek's subjects invented convoluted meteorological theories to interpret data suggesting that temperature *increased* as one got further north of the equator.

Conclusions

People are constantly forming, testing, and revising hypotheses about how the world works, what will happen next, or what the consequences of an action will be. This informal scientific reasoning differs in important ways from the prototype of science as the discovery of laws. Theories about everyday phenomena are of limited explanatory power and scope, and data are prone to considerable error and ambiguity. As research on cue learning illustrates, uncertainty and complexity in the environment affect the nature of hypotheses, the goals of hypothesis testing, and the processes by which data are encoded, aggregated, and interpreted.

Of course, uncertainty and complexity are encountered in the formal practice of science as well. However, the professional scientist is in a position to use tools such as controlled experimentation and statistical methods. These techniques certainly help with the problems of testing and revising hypotheses in a probabilistic environment. Even informal settings sometimes permit experimentation and quantification. Educated laypeople have a fair degree of intuition about some basic principles of experimentation, and they may learn and

reason more effectively when they are able to apply those principles (Klayman, 1988; Nisbett et al., 1983).

On the other hand, the availability of scientific methods does not eliminate the difficulties of using ambiguous data to test and revise hypotheses. The conduct of science is seldom as neat and clear as the law-discovery prototype, or as the resulting published articles make it sound (see, e.g., Mitroff, 1974). Informal thinking plays an important role in forming new theories and hypotheses, choosing and evaluating research methods, and interpreting findings. Thus, many of the phenomena of informal reasoning are likely to have relevance to the professional conduct of science as well, especially in less exact domains.

A thorough understanding of informal scientific reasoning is still a long way off. However, there is some convergence in recent work using a variety of different paradigms: Hypothesis testing is being viewed in a broader context, as one of the critical steps in an interactive process of discovery, testing, and revision of ideas (Holland, et al., 1986; Johnson-Laird, 1983; Klahr & Dunbar, 1988; Klayman & Ha, 1987; Lakatos, 1978; Langley et al., 1987). The course of this reasoning process is a function of world knowledge, prior theories, basic processing characteristics, and information from data. This view of hypothesis testing is much more complex than earlier models, but it is also more likely to be at least half right.

References

Adelman, L. (1981). The influence of formal, substantive, and contextual task properties on the relative effectiveness of different forms of feedback in multiple-cue probability learning tasks. *Organizational Behavior and Human Performance*, *27*, 423-442.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*, 1-27.

Brehmer, B. (1980). In one word: not from experience. *Acta Psychologica*, *45*, 223-241.

Brehmer, B. (1987). Note on subjects' hypotheses in multiple-cue probability learning. *Organizational Behavior and Human Decision Processes*, *40*, 323-329.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd Ed.). Berkeley: University of California Press.

Camerer, C. (1981). *The validity and utility of expert judgment.* Unpublished doctoral dissertation, University of Chicago, Graduate School of Business.

Einhorn, H.J. & Hogarth, R.M. (1986). Judging probable cause. *Psychological Bulletin*, *99*, 3-19.

Fischhoff, B. & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*, 239-260.

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.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery.* Cambridge, MA: MIT Press.

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge, MA: Harvard University Press.

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

Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 317-330.

Klayman, J. (in press). On the how and why (not) of learning from outcomes. In B. Brehmer & C. R. B. Joyce (Eds.), *Human Judgment: The Social Judgment Theory Approach.* Amsterdam: North-Holland.

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

Klayman, J. & Ha, Y. (1988). *Hypothesis testing in rule discovery: Strategy and structure* (Working Paper No.133). Chicago: University of Chicago, Graduate School of Business, Center for Decision Research.

Lakatos, I. (1978). *The methodology of scientific research programmes*. London: Cambridge University Press.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32, 311-328.

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

Miller, P.McC. (1971). Do labels mislead? A multiple cue study, within the framework of Brunswick's probabilistic functionalism. *Organizational Behavior and Human Performance, 6*, 480-500.

Mitroff, I. (1974). *The subjective side of science*. Amsterdam: Elsevier.

Muchinsky, P.M. & Dudycha, A.L. (1975). Human inference behavior in abstract and meaningful environments. *Organizational Behavior and Human Performance, 13*, 377-391.

Mynatt, C.R., Doherty, M.E. & Tweney, R.D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology, 30*, 395-406.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90*, 339-363.

Nisbett, R. E. & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Ross, L. & Lepper, M.R. (1980). The perseverance of beliefs: Empirical and normative considerations. In R.A. Shweder (Ed.), *Fallible Judgment in Behavioral Research: New Directions for Methodology of Social and Behavioral Science* (Vol 4, pp. 17-36). San Francisco: Jossey-Bass.

Sniezek, J.A. (1986). The role of variable cue labels in cue probability learning tasks. *Organizational Behavior and Human Decision Processes, 38*, 141-161.

Tweney, R. D. (1985). Faraday's discovery of induction: A cognitive approach. In D. Gooding & F. James (Eds.), *Faraday rediscovered* (pp. 159-209). London: MacMillan.

Wason, P.C. & Johnson-Laird, P.N. (1972). *Psychology of Reasoning: Structure and Content*. London: Batsford.

Preparation of this article was supported by grant SES-8706101 from the Decision, Risk, and Management Science program of the National Science Foundation. Thanks to Jackie Gnepp and to my colleagues at the Center for Decision Research for their helpful comments.