Blackwell Handbook of Judgment and Decision Making

Edited by

Derek J. Koehler and Nigel Harvey

10

Hypothesis Testing and Evaluation

Craig R. M. McKenzie

Introduction

Imagine a physician examining a patient who exhibits certain symptoms. The physician would undoubtedly generate possible explanations, or hypotheses, regarding the cause of the symptoms. If additional evidence were needed to confidently diagnose the patient, the physician might need to decide which questions to ask the patient or which tests to order. Then, based on the answers to the questions or the results of the tests, the physician would update confidence in one or more hypotheses and perhaps feel confident enough to recommend a course of action. If still more evidence were needed – perhaps the test results were unexpected, leaving no obvious explanation of the patient's symptoms – decisions would have to be made again about how to gather more information.

The above scenario captures the essence of hypothesis development (Klayman, 1995), which is generally concerned with how we determine whether there is a match between what we think might be true about the world and what is in fact true about the world. The process is a complex set of behaviors, but for present purposes, it will be seen as consisting of three stages: Hypothesis generation, testing, and evaluation. In the above example, hypothesis *generation* occurs when the physician produces at least one hypothesis based on the patient's symptoms. What is (are) the most likely explanation(s) of the pattern of symptoms? Once the physician has a hypothesis, but is unsure whether it is correct, more information might be collected for *testing* the hypothesis: Which medical tests should be run, which questions should be asked? Once the hypothesis has been put to test, the results are used to *evaluate* the hypothesis. Do the results provide confirming or disconfirming evidence (or neither)? How strongly do the results (dis)confirm the hypothesis?

Hypothesis development is not limited to formal situations such as physicians diagnosing patients and scientists testing theories. More mundane examples include determining whether your new research assistant is reliable, or how to get a young child to eat

more vegetables. As will be argued shortly, hypothesis development is probably even more mundane than you think: We constantly engage in it in order to make sense out of our infinitely complex environment.

Due to its ubiquity, complexity, and importance, hypothesis development has been, and continues to be, the focus of much psychological research. The purpose of this chapter is to provide an overview of some important issues in the psychology of hypothesis development. As the title suggests, the emphasis is on the second and third stages, namely, hypothesis testing and evaluation. The first section notes some key findings beginning in the 1950s regarding the importance of hypothesis development. The second section critically examines "confirmation bias," a term now used to describe people's purported tendency to be overly confident in their favored hypothesis. The third section reviews recent research indicating that consideration of the environment outside the laboratory is crucial for understanding hypothesis-testing behavior inside the laboratory. The chapter concludes with an overview of the main points and their implications for understanding how, and how well, people put their ideas to test.

Some Early Findings

Bruner, Goodnow, and Austin's (1956) book represents a milestone in research on hypothesis development, even though it was only indirectly concerned with the topic. Its primary focus was on "concept attainment," or how we learn to group objects into pre-existing categories. In a typical experiment, participants were presented with cards showing different shapes that varied in terms of size and color. There was a predetermined "concept" and participants were to learn, by choosing cards, which attributes distinguished exemplars from non-exemplars. The experiment usually began with the experimenter providing the participant with an exemplar of the concept to be learned. After each card was chosen by the participant, the experimenter revealed whether it was an instance of the concept. For example, the concept might be "triangle," in which case every chosen card with a triangle would receive a positive response from the experimenter and all others a negative response. In this way, the participant would learn to attend to shape and to ignore color and size.

For our purposes, what is most important about the research of Bruner et al. (1956) is how it paved the way for replacing the (behaviorists') passive view of the mind with one in which people actively organized their experience and brought their knowledge to bear on the task. In this case, participants often *actively engaged in hypothesis development* regarding the correct concept. For example, after being shown that a large red triangle was an instance of the concept, a participant might hypothesize that "large shapes" was the correct concept. This would then influence not only the next card selected, but also the interpretation of the subsequent feedback (depending on the specific strategy used). A different participant might hypothesize that "red triangles" was the correct concept and therefore behave differently. The insight that participants were actively testing hypotheses was crucial for understanding behavior – both success and failure – in these concept attainment tasks.

Other learning phenomena were later uncovered that both confirmed and extended Bruner et al.'s (1956) active hypothesis-development viewpoint. For example, non-learning sometimes occurred for presumably smart and motivated participants (e.g., Levine, 1971). Some participants simply failed to learn the concept or rule despite many trials with feedback. This is problematic for the passive view of the mind in which learning is gradual and automatic, but it makes sense from a hypothesis-development viewpoint: Non-learning occurred when participants failed to generate the correct hypothesis. Closely related was the finding that some learning tended to be all-or-none rather than gradual (e.g., Restle, 1965). Participants' performance was often at chance level until they generated the correct hypothesis and was virtually perfect thereafter.

Also related was the finding that organisms do not learn to make a connection between all events equally well. For example, in a classic study, Garcia and Koelling (1966) allowed thirsty rats to drink flavored water while simultaneously presented with a light and noise. Some rats were then mildly poisoned while others received a shock. Those who received the poison showed aversion to the flavor of the water, but not to the light and noise, while those who received the shock showed an aversion to the light and noise, but not to the flavor of the water. These findings indicate that certain hypotheses are more readily generated, or seen as more plausible a priori, in certain contexts. Again, a passive view of the mind cannot explain this.

All of the above research examined learning using discrete variables (e.g., shapes were either triangles or squares, red or blue). Interestingly, however, similar results have been found in tasks examining how participants learn to predict scaled criterion values based on scaled cue values. These cue-learning tasks are also solved through the process of hypothesis development. In this case, participants test a series of hypotheses about the possible functional form, and the hypotheses are tested in a systematic order (Brehmer, 1974; Sniezek & Naylor, 1978; for an overview, see Brehmer, 1980). Participants first check to see if the relationship is linear and positive. If this turns out to be incorrect, they test the hypothesis that the relationship is linear and negative, followed by an inverse U-shaped hypothesis, then a U-shaped hypothesis, and so on. The amount of time taken to discover the true relationship depends on how far down the participant's hypothesis hierarchy the true rule is.

The above findings indicate that it is simply not the case that hypothesis development is primarily the province of scientists testing theories and physicians diagnosing illnesses. We all engage in hypothesis development on a regular basis as we try to organize and impose structure on our complex world.

Given its prevalence and importance, how good are we at developing our hypotheses? Bruner et al. (1956) found that participants did not behave optimally. Because of the circumscribed (and usually deterministic) nature of the tasks, a participant could, in theory, rule out multiple hypotheses simultaneously with a single test. Instead, though, participants tended to use a win-stay, lose-shift strategy. If a test of the current hypothesis led to the expected positive response, participants would continue testing it. If a test led to an unexpected negative response, only then would they consider a different hypothesis. Thus, participants tended to test hypotheses in a serial fashion rather than in parallel, which led to inefficiencies; participants were not getting as much information out of each test as they could have. Nonetheless, it is worth noting that Bruner et al.

were impressed by participants' sensitivity to task demands. Participants tended to use more sophisticated strategies as task demands were reduced. Indeed, those who tried to use complex strategies when task demands were high sometimes performed worse than those using simpler strategies (for similar findings in the context of choice strategies, see Payne, Bettman, & Johnson, 1993).

In sum, hypothesis development is a ubiquitous means for coping with a complex world. Hypotheses enable us to interpret incoming data by telling us what to look for; they suggest what's relevant. Indeed, we must have some idea of what might be learned before we can learn it.

Confirmation Bias

Wason (1960) was not as sanguine as Bruner et al. (1956) with respect to people's hypothesis-development abilities. In particular, Wason viewed the common strategy of testing hypotheses serially, only revising hypotheses after receiving disconfirming evidence (Bruner et al.'s "successive scanning"), as being seriously deficient. He devised the "2-4-6" concept attainment task in order to demonstrate this.

In this task, participants were told that the experimenter had a rule in mind that produces triples of numbers, an example of which was 2-4-6. They were to produce triples of numbers in order to figure out the experimenter's rule. After announcing each triple, participants were told whether or not it conformed to the experimenter's rule. They could test as many triples as they wished and were to state what they thought was the correct rule only after they were highly confident they had found it. Few participants discovered the correct rule (with their first "highly confident" announcement), which was "numbers in increasing order of magnitude."

How could most participants be so confident in a wrong rule after being allowed to test it as much as they wished? The initial 2-4-6 example naturally suggests a hypothesis such as "increasing intervals of two" (which was the most commonly stated incorrect rule). They would then test their hypothesis by stating triples such as 8-10-12, 14-16-18, 20-22-24, and 1-3-5 – triples that were consistent with their hypothesized rule. Of course, each of these triples is consistent with the correct rule as well, and hence participants received a positive response from the experimenter ("Yes, it conforms to the rule"). This, in turn, led them to believe incorrectly that they had discovered the correct rule.

Wason (1960) claimed that participants appeared unwilling to test their hypotheses in a manner that would lead them to be disconfirmed (which is what Popper, 1959, claimed was how people ought to test hypotheses). The only way to disconfirm the "increasing intervals of two" hypothesis is to test triples that are *not* expected to conform to the hypothesis, such as 2-4-7 or 1-2-3 (Bruner et al. referred to these as "indirect tests"). Testing such triples would lead to unexpected "yes" responses from the experimenter, and participants would then (and only then) know that their hypothesis was wrong. Instead, Wason argued, participants tested their hypotheses in a way that led them to be confirmed. This came to be known as "confirmation bias" and created a stir because of the apparent dire implications: We test our hypotheses in a manner that leads

us to believe them, regardless of their correctness. This view of lay hypothesis development became dominant in psychology (Evans, 1989; Mynatt, Doherty, & Tweney, 1977, 1978; Nisbett & Ross, 1980; Snyder, 1981; Snyder & Campbell, 1980; Snyder & Swann, 1978). As Nickerson (1998, p. 175) recently noted, "If one were to attempt to identify a single problematic aspect of human reasoning that deserves attention above all others, the confirmation bias would have to be among the candidates for consideration."

But what exactly is confirmation bias? The label has been applied to a variety of phenomena (Fischhoff & Beyth-Marom, 1983; Klayman, 1995; Klayman & Ha, 1987; Nickerson, 1998; Poletiek, 2001). For present purposes, I will refer to confirmation bias as testing or evaluating a hypothesis such that inappropriately high confidence in the hypothesis is the systematic result (Klayman, 1995; Nickerson, 1998). Accordingly, I briefly review findings from hypothesis-testing and hypothesis-evaluation tasks separately below.

Testing strategies and confirmation bias

Despite the enormous implications of confirmation bias, it has become clear that many of the early claims were overstated. For example, it is now generally accepted that hypothesis-testing strategies do not, by themselves, necessitate confirmation bias (Klayman, 1995; Poletiek, 2001). In order to discuss the testing stage of hypothesis development and how it might relate to confirmation bias, I will move away from Wason's 2-4-6 task, which involves hypothesis generation, testing, *and* evaluation. (I will return to the task in the section on "Importance of the environment.") Using a simple probabilistic testing task, I will first discuss how one *ought* to select among possible tests in order to determine, in the most efficient fashion, which of the competing hypotheses is most likely true. Subsequently, I will discuss how people choose tests in such a task, and what implications, if any, this has for confirmation bias.

Consider Table 10.1, which lists the proportion of whimsical creatures, Gloms and Fizos, on the planet Vuma, possessing each of eight features (Skov & Sherman, 1986; see also Slowiaczek, Klayman, Sherman, & Skov, 1992). For example, the first feature

Table 10.1 Percentage of Gloms and Fizos possessing each of eight features

Feature	*Gloms (%)*	*Fizos (%)*
1 – wear hula hoops	10	50
2 – eat iron ore	28	32
3 – have gills	68	72
4 – gurgle a lot	90	50
5 – play the harmonica	72	68
6 – drink gasoline	32	28
7 – smoke maple leaves	50	90
8 – exhale fire	50	10

regards whether the creatures wear hula hoops; 10 percent of Gloms do so and 50 percent of Fizos do so. Assume that there are equal numbers of Gloms and Fizos on the planet and there are only these two types of creature. If you had to ask a randomly sampled creature from Vuma about just one of the listed features in order to determine whether it was a Glom or a Fizo, which feature would you choose?

Deciding which question one ought to ask requires a fair amount of sophistication, even in this highly constrained and well-specified example. A complicating factor is that you don't know what answer the creature will give. We need to first think about the possible answers to a question (e.g., "Yes, I wear a hula hoop" or "No, I don't wear a hula hoop") in terms of their likelihood ratios, $p(D/H1)/p(D/H2)$, where D corresponds to data and $H1$ and $H2$ correspond to the competing hypotheses. In this context, the datum is the answer to the question ("yes" or "no") and the hypotheses are that the creature is a Glom ($H1$) or a Fizo ($H2$). To the extent that the likelihood ratio differs from 1, the datum is diagnostic, or helps discriminate between the hypotheses. For feature 1 in Table 10.1, the likelihood ratio for a "yes" answer is 0.1/0.5 and for a "no" answer is 0.9/0.5. Note that the first ratio is less than one and the second is greater than one. This is because the "yes" answer provides evidence against $H1$ and the "no" answer provides evidence for $H1$.

How confident should one be in $H1$ after receiving one of these answers? Bayes' theorem provides an answer. Generally:

$$p(H1/D) = p(H1)p(D/H1)/[p(H1)p(D/H1) + p(H2)p(D/H2)]. \quad (10.1)$$

So, in the case of a "yes" response to a question about feature 1 ($F1$), confidence that the creature is a Glom should be:

$$p(\text{Glom}/F1) = 0.5(0.1)/[0.5(0.1) + 0.5(0.5)] = 0.17.$$

However, if the creature answers "no" ($\sim F1$), confidence in the Glom hypothesis should be:

$$p(\text{Glom}/\sim F1) = 0.5(0.9)/[0.5(0.9) + 0.5(0.5)] = 0.64.$$

Note that, relative to the prior confidence in the Glom hypothesis (0.5), the "yes" answer decreased confidence more than the "no" answer increased confidence because the former is more diagnostic than the latter. The former likelihood ratio is 1/5 and the latter is 9/5. One way to compare the diagnosticity of two data, when one likelihood ratio is greater than 1 and the other is less than 1, is to take the reciprocal of the ratio that is less than 1 and then make a direct comparison. In this case, $5/1 > 9/5$. Another way is to take the absolute value of the log of each ratio (log likelihood ratio, or LLR):

$$\text{abs}[\log_2(1/5)] = 2.3 > \text{abs}[\log_2(9/5)] = 0.85.$$

Given that the different answers to the same question will be differentially diagnostic (except when the two likelihoods sum to 1), and the tester does not know which answer the creature will give, how should one choose a question? One way is to select the

question with the highest *expected* LLR. In order to do so, one must calculate, for each question, the LLR of each possible answer ("yes" or "no") and how likely each answer is, which depends on both the frequency of the feature for each creature and how likely each hypothesis is. In the case of feature 1, you would expect to hear "yes" 30 percent of the time because 50 percent of the creatures are Gloms, 10 percent of whom will answer yes, and 50 percent of the creatures are Fizos, 50 percent of whom will answer "yes." Similarly, you expect to hear "no" 70 percent of the time. Thus, for feature 1, the expected (absolute) LLR is $0.3(2.3) + 0.7(0.85) = 1.285$. Now just make such calculations for the remaining seven features so you can choose the feature with the highest expected LLR. Simple, right?

It's fair to say that choosing the most diagnostic question is not so simple, even in this simple example. How, then, do people choose which questions to ask? Given a set of questions as in Table 10.1, three factors seem to drive people's choices: Diagnosticity, positivity, and extremity. Diagnosticity essentially refers to expected LLR and has been found to be a major determinant of question selection (Bassok & Trope, 1984; Skov & Sherman, 1986; Slowiaczek et al., 1992; Trope & Bassok, 1982, 1983). In other words, one strong tendency is to select the questions that ought to be selected in order to test hypotheses most efficiently. In Table 10.1, features 1, 4, 7, and 8 have the highest expected LLR. Participants might tend to choose questions with the highest expected LLR because the measure is highly correlated with the algebraic difference between the two likelihoods (when the prior probabilities for the hypotheses are equal; Slowiaczek et al., 1992). That is, to the extent that the difference between the two percentages listed for a given feature in Table 10.1 is large, expected LLR is large. Participants might be sensitive to this difference, which correlates with expected LLR.

The second factor, positivity, is the tendency to ask questions that are expected to result in a "yes" response, given the truth of the working hypothesis (Klayman & Ha, 1987; Skov & Sherman, 1986; Bruner et al., 1956, referred to these as "direct tests"). This is analogous to the testing strategy that participants in Wason's (1960) 2-4-6 task appeared to use and is now commonly referred to as "positive hypothesis testing" (Klayman & Ha, 1987). For example, if testing the Glom hypothesis and forced to choose between asking about features 1 or 4 (which have the same expected LLR), participants tend to prefer feature 4.

Extremity, the final factor, refers to a preference for questions whose outcomes are very likely or unlikely under the working hypothesis relative to the alternate hypothesis. Thus, if testing the Glom hypothesis, participants prefer asking about feature 4 over feature 8, though asking about the features has the same expected LLR and both are positive tests.

These latter two tendencies look like they might lead to inappropriately high confidence in the hypothesis being tested, or confirmation bias. But they don't. At least not necessarily. Any testing strategy not solely concerned with diagnosticity will be inefficient (assuming equal costs of information and errors), but as long as the tester takes into account the test biases at the subsequent hypothesis-evaluation stage, no confirmation bias will result. As Klayman (1995) noted, participants in Wason's (1960) experiment erred not in their tendency to conduct positive tests (i.e., test triples that they expected

to result in "yes" answers), but in failing to take this into account at the hypothesis-evaluation stage. In this case, they failed to notice that their strategy left open the possibility of false negative errors. Similarly, if one evaluates the hypothesis properly (Equation 10.1) following "extreme" questions, there will be no bias.

Even other forms of biased testing, such as a tendency to recruit facts or arguments to support, rather than refute, the working hypothesis do not necessitate confirmation bias if one takes the testing bias into account at the evaluation stage (Klayman, 1995). For example, one might be sensitive to how difficult it is to recruit positive evidence, or not be influenced much by the positive evidence, or be strongly influenced by even the slightest negative evidence (e.g., McKenzie, Lee, & Chen, 2002). In short, biased testing strategies lead to inefficiencies but do not necessitate confirmation bias.

Evaluation strategies and confirmation bias

While there is a huge literature on how people evaluate hypotheses after finding out the result of a test, I'll just mention a couple of phenomena that seem most relevant to confirmation bias (see Klayman, 1995, and Nickerson, 1998, for other examples). Perhaps the most obvious one is that people tend to be more skeptical of new information that is inconsistent with their favored hypothesis than consistent with it (Koehler, 1993; Lord, Ross, & Lepper, 1979). That is, participants sometimes downplay evidence against their favored hypothesis. Although this would appear to lead to confirmation bias, it can be normatively reasonable for prior beliefs to influence perceived evidence quality (Koehler, 1993; Lord et al., 1979). For instance, if someone were to tell you that Earth was cone-shaped, should you decrease confidence in your current belief about Earth's shape or dismiss the new "data"? The extent to which evidence inconsistent with a favored hypothesis ought to result in a change in confidence rather than be met with skepticism is a difficult normative issue. When both data and hypotheses are uncertain, the relationship between them is mutual; each can inform the other (Thagard, 1989). (Interestingly, the first significance tests were used to reject data [outliers], not hypotheses; Gigerenzer, Swijtink, Porter, Daston, Beatty, & Krüger, 1989, pp. 80–4.) It is of course possible that people are overly eager to dismiss evidence inconsistent with their beliefs, but without a clear normative benchmark, we cannot know for sure.

In addition, people sometimes interpret ambiguous evidence in ways that give the benefit of the doubt to their favored hypothesis. Whether you interpret someone's failure to return a smile from across the room as indicating the person didn't see you or the person is snubbing you will likely be influenced by whether the person is a good friend or is known for being socially distant. As Nisbett and Ross (1980) point out, however, this is not necessarily an error. Under these circumstances, it generally would be more likely true that the person didn't see you if he or she were a good friend, and it generally would be more likely true that the person was snubbing you if he or she were socially distant. It might very well be that people interpret ambiguous data in overly generous ways, but the complexity of the normative issue makes it difficult to know if or when people are making errors (Klayman, 1995).

Confirmation bias due to interactions between testing and evaluation

Although neither testing strategies nor evaluation strategies, by themselves, appear to lead to confirmation bias, some combinations of the two can (Klayman, 1995; Poletiek, 2001; Slowiaczek et al., 1992). Klayman (1995) notes three such combinations. First, confirmation bias can result from the combination of positive testing (in this case, asking questions that you expect a "yes" answer to, if your hypothesis is correct) and the fact that respondents are biased to answer "yes" to questions in social settings ("acquiescence bias"; Zuckerman, Knee, Hodgins, & Miyake, 1995). If interviewers do not take respondents' biases into account – and apparently they don't – this leads to higher confidence in the working hypothesis than is warranted. Though interesting, note that this example of confirmation bias is limited to asking yes/no questions in social settings.

Two other examples appear to have broader applicability. One is the combination of positive testing – in this case, asking about features expected to be present if the working hypothesis is true – and the fact that participants are more affected by the presence of features than by their absence (e.g., feature-positive effects; Jenkins & Sainsbury, 1969, 1970; Newman, Wolff, & Hearst, 1980). Because positive testing implies that the presence of features confirms the hypothesis and their absence disconfirms the hypothesis, and feature-positive effects imply that presence has more impact than absence, then evidence favoring the hypothesis will have the most impact.

The third and final combination noted by Klayman (1995) is that of preferring extremity and finding confirming and disconfirming outcomes more equal in terms of their informativeness than they really are (Slowiaczek et al., 1992). In terms of Table 10.1, the preference for extremity implies asking about features 1 and 4 if Glom were the working hypothesis (because for each feature the likelihood corresponding to Gloms is much closer to 0 or 1 than the corresponding likelihood for Fizos). Note that the confirming answer is "no" after asking about feature 1 and "yes" after asking about feature 4. Recall that confirming and disconfirming test outcomes are often differentially informative. For example, when testing the Glom hypothesis and asking about feature 1, it was shown earlier that the confirming "no" answer should increase confidence from 0.5 to 0.64, whereas the disconfirming "yes" answer should decrease confidence from 0.5 to 0.17. The latter change is larger because the disconfirming "yes" answer is more diagnostic than the confirming "no" answer. For both features 1 and 4, the confirming outcomes have likelihood ratios of 9/5 whereas the disconfirming outcomes have likelihood ratios of 5/1. Participants, however, tend to see the different test outcomes as more similar in terms of diagnosticity than they ought to. For example, Slowiaczek et al. (1992) found that participants reported confidence in the working Glom hypothesis of 0.62 and 0.27 to "no" and "yes" answers, respectively, for likelihoods analogous to feature 1 in Table 10.1. Note that the former confidence report is slightly too low, but that the latter is considerably too high, giving an overall edge to the working hypothesis. Generally, when selecting features that are more extreme under the working hypothesis, the confirming outcome is less diagnostic than the disconfirming outcome. (The reason, discussed in the next section, is that the disconfirming outcome under these conditions is rarer, or more surprising.) Because the weakly confirming and strongly disconfirming

outcomes will be seen as being more similar in terms of their diagnosticity than they really are, confidence in the working hypothesis will tend to be higher than is warranted.

Confirmation bias summary

Early claims about confirmation bias appear to have been overstated. It is now generally accepted that neither hypothesis-testing strategies nor hypothesis-evaluation strategies, by themselves, appear to lead to confirmation bias, but working together they can (Klayman, 1995; Poletiek, 2001; Slowiaczek et al. 1992).

Importance of the Environment in General, and Rarity in Particular

The foregoing discussion was concerned with the typical formalizations associated with hypothesis testing and evaluation (e.g., likelihoods) that are often manipulated orthogonally in the laboratory. Sometimes overlooked in such analyses, however, is what the situation tends to be like outside the laboratory and how these "real world" conditions might influence behavior inside the laboratory. In this section, recent research on testing and evaluation behavior that has been influenced by considerations of the environment is discussed. In particular, the focus is on the *rarity of data*. How rare, or surprising, data are is a key notion in essentially all formal theories of hypothesis testing (Poletiek, 2001, Chapter 2) and it will be argued that people are highly sensitive to this variable. Indeed, even in tasks in which considerations of rarity might seem irrelevant, participants nonetheless appear to make (reasonable) assumptions about rarity based on experience outside the laboratory, which can lead their behavior to be seen as less sensible than it really is.

Wason's "2-4-6" task

Klayman and Ha's (1987) analysis of Wason's "2-4-6" task illustrates the importance of taking into account environmental conditions in general, and rarity in particular, when trying to understand hypothesis-development behavior. Recall that Wason (1960) argued that people were prone to confirming their hypotheses because they tested triples they expected to lead to a "yes" response from the experimenter (positive hypothesis tests). Klayman and Ha pointed out that Popper (1959), whose view of hypothesis testing Wason considered normative, had prescribed testing hypotheses so that they are most likely to be disconfirmed; he did not say that one ought to test cases that the hypothesis predicts will fail to occur. In other words, Klayman and Ha distinguished between disconfirmation as a goal (as prescribed by Popper) and disconfirmation as a testing strategy. Wason (1960) confounded these two notions. Because the true rule ("increasing numbers") is more general than the tentative "increasing intervals of two"

hypothesis, the only way to disconfirm the latter is by testing triples that are expected not to work (negative hypothesis tests). This, of course, is just what Bruner et al. (1956) found people tend not to do, which is why Wason designed his task as he did. But notice that the situation could easily be reversed: One could entertain a hypothesis that is more general than the true rule, in which case the only way to disconfirm the hypothesis is by testing cases hypothesized to work (and finding they do not) – exactly opposite from the situation in Wason's task. For example, an advertising executive might hypothesize that advertising a particular product on television is the key to success, whereas in fact only advertising on prime time television will work. In this situation, testing only cases hypothesized *not* to work (e.g., advertising on the radio) could lead to incorrectly believing the hypothesis (because all the cases that the hypothesis predicts will not work will, in fact, not work).

Whether positive testing is a good strategy, then, depends on the relationship between the hypothesized and true rule. Furthermore, positive testing is more likely than negative testing to lead to disconfirmation when (a) you are trying to predict a rare event, and (b) your hypothesized rule includes about as many cases as the true rule does (i.e., your hypothesis describes an equally rare event). Finally – and very important – the above two conditions (both involving rarity), Klayman and Ha (1987) argue, are commonly met in real-world hypothesis-testing situations, implying that positive hypothesis testing is generally more likely than negative hypothesis testing to lead to disconfirmation.

Thus, despite the results from Wason's (1960) 2-4-6 task, positive testing appears to be a highly adaptive strategy for testing hypotheses under typical real-world conditions. This virtual reversal of the perceived status of testing cases expected to work is primarily due to Klayman and Ha's *analysis of the task environment.* Seen independent of the environmental context in which it is usually used, positive testing can look foolish (as in Wason's task). Seen in its usual environmental context, it makes good normative sense. Klayman and Ha's work underscores the point that understanding hypothesis-development behavior requires understanding the context in which it usually occurs.

Wason's selection task

Wason is also well known for a second hypothesis-testing task: the selection task (Wason, 1966, 1968). In this task, which involves only testing (and not hypothesis generation or evaluation), participants have four cards in front of them. Imagine, for example, that each has a letter on one side and a number on the other. The visible side of one card has an "A" on it, a second has a "K", a third a "2", and the fourth a "7". You are to test whether the following rule is true or false: If a card has a vowel on one side, it has an even number on the other. Which cards must you turn over in order to test the rule?

According to one interpretation of the rule ("material implication"), propositional logic dictates that the A and 7 cards should be turned over. When testing "If P, then Q" ($P \rightarrow Q$), only the combination of P and not-Q (a vowel on one side and an odd number on the other in the example) falsifies the rule; any other combination is consistent with the rule. Thus, turning over the "A" card is useful because an even number on the other side is consistent with the rule, but an odd number is not. Similarly, turning

over the "7" card is useful because finding a consonant on the other side is consistent with the rule, but finding a vowel is not. By contrast, nothing can be learned by turning over the "K" and "2" cards because whatever is on the other side of either card is consistent with the rule. Therefore, there is no point in turning over these two cards according to propositional logic.

Which cards do participants select? The most common response is to select the "A" and "2" cards (P and Q). Typically, fewer than 10 percent of participants request the logically correct "A" and "7" (P and not-Q) combination (Wason, 1966, 1968). Participants tend to select the cards mentioned in the rule. This has traditionally been seen as a classic demonstration of irrationality.

However, Oaksford and Chater (1994; see also Nickerson, 1996) have shown that the P and Q cards are the most informative if one assumes (a) an inferential (Bayesian) approach to the task that treats the cards to be turned over as a sample from a larger population of interest, and (b) that P and Q are rare relative to not-P and not-Q (the "rarity assumption"). Their model enabled them to account for a wide variety of selection task findings and, equally important, led to predictions as to when participants would prefer to turn over the not-Q card. Indeed, as predicted, participants are more likely to turn over the not-Q card as P and Q become more common (Oaksford & Chater, 2003).

Thus, Oaksford and Chater (1994) make two assumptions that appear to reflect the real world. First, they assume an inferential approach that is appropriate in a probabilistic, rather than deterministic, environment. Second, they assume (and supporting empirical evidence is discussed below) that conditional rules or hypotheses tend to mention rare, not common, events, which plausibly describes everyday discourse. Normative principles, combined with considerations of the environment, can help explain behavior in the selection task.

Impact of confirming evidence

In most of the discussion thus far, there have been two kinds of confirming evidence: outcomes expected to occur that do occur and outcomes expected not to occur that don't. Judging by their preferred testing strategies, participants consider the former confirming outcomes more informative than the latter. For example, people are much more likely to perform positive hypothesis tests than negative hypothesis tests, suggesting that people are generally more concerned about whether what they expect to occur does occur than whether what they expect not to occur doesn't.

Generally, when evaluating hypotheses of the form P $\rightarrow$ Q, participants deem the combination of P and Q (P&Q outcomes) as more informative than not-P&not-Q outcomes, although both provide confirming evidence. Why? One possible reason comes from Oaksford and Chater's (1994) rarity assumption: When testing P $\rightarrow$ Q, maybe P and Q tend to be rare relative to not-P and not-Q. If so, then from a Bayesian perspective, the P&Q outcome *is* more informative than a not-P&not-Q outcome.

To see why a combination of rare events is most informative, consider testing a forecaster's claim of being able to predict the weather in San Diego, where rainy days are

rare. Assume that the forecaster rarely predicts rain and usually predicts sunshine. On the first day, the forecaster predicts sunshine and is correct. On the second day, the forecaster predicts rain and is correct. Which of these two correct predictions would leave you more convinced that the forecaster can accurately predict the weather and is not merely guessing? The more informative of the two observations is the correct prediction of rain, the rare event, at least according to Bayesian statistics (Horwich, 1982; Howson & Urbach, 1989). Qualitatively, the reason is that it would not be surprising to correctly predict a sunny day by chance in San Diego because almost every day is sunny. That is, even if the forecaster knew only that San Diego is sunny, you would expect her to make lots of correct predictions of sunshine just by chance alone. Thus, such an observation does not help much in distinguishing between a knowledgeable forecaster and one who is merely guessing. In contrast, because rainy days are rare, a correct prediction of rain is unlikely to occur by chance alone and therefore provides relatively strong evidence that the forecaster is doing better than merely guessing. Generally, given two dichotomous variables, P and Q, when testing $P \rightarrow Q$, a P&Q outcome will be more informative than a not-P&not-Q outcome whenever $p(P) < 1 - p(Q)$ (Horwich, 1982; Mackie, 1963). This condition is clearly met when both P and Q are rare ($p < 0.5$).

Thus, just as with the above account of Wason's selection task, it is possible to explain a preference for the confirming outcome mentioned in the hypothesis by adopting a Bayesian approach and by making the rarity assumption. Furthermore, the account leads to the prediction that this preference will be reversed when it is clear to participants that the hypothesis mentions common, not rare, events.

McKenzie and Mikkelsen (2000) tested this prediction. They found that participants evaluating abstract hypotheses tended to consider whichever combination of events was mentioned in the hypothesis to provide the strongest support. However, participants evaluating concrete hypotheses about variables they were familiar with tended to select the combination of rare events as most informative, regardless of whether the events were mentioned in the hypothesis. In other words, when the hypothesis mentioned common events, participants tended to consider the *un*mentioned confirming observation as providing the strongest support.

These results suggest that participants generally assume that mentioned observations are rare when evaluating abstract, unfamiliar hypotheses (the norm in the laboratory), but this default assumption is overridden when it is clear that it does not apply. Once again, participants' behavior is consistent with the qualitative use of Bayesian principles combined with reasonable assumptions about how the world usually works.

Direct evidence for the rarity assumption

The rarity assumption has been important for explaining, in qualitatively normative terms, purported errors in the hypothesis-development tasks discussed above. That is, if it is assumed that hypotheses tend to mention rare events, much explanatory power is gained. In fact, several authors have speculated that people do indeed tend to hypothesize about rare events (e.g., Einhorn & Hogarth, 1986; Klayman & Ha, 1987; Mackie,

1974; Oaksford & Chater, 1994). Rare or unexpected events "demand" explanation. We tend to hypothesize about the factors leading to success, not mediocrity; about the factors leading to being HIV+, not HIV−; about what caused a plane crash, not a normal flight; and so on. Consistent with these speculations, McKenzie, Ferreira, Mikkelsen, McDermott, & Skrable (2001) found that participants had a tendency – often a strong one – to phrase conditional hypotheses ("If _____, then _____") in terms of rare rather than common events. This provides an answer to the question of why, as a default strategy, people consider mentioned confirming observations to be more informative than unmentioned confirming observations: mentioned observations generally *are* more informative because they are rare.

Covariation assessment

Research on covariation assessment is concerned with evidence evaluation, and participants' assumptions about, and sensitivity to, the rarity of data can explain purported errors in this highly studied task as well. In a typical task, participants are asked to assess the relationship between two variables, each of which can be either present or absent, resulting in the familiar 2 × 2 matrix (see Table 11.1 in the next chapter). Cell A corresponds to the joint presence of the variables, Cell B to the presence of variable 1 and the absence of variable 2, Cell C to the absence of variable 1 and the presence of variable 2, and Cell D to the joint absence of the variables. Given the four cell frequencies, participants assess the direction or strength of the relationship (for reviews, see Allan, 1993; McKenzie, 1994; Chapter 11, this volume). Assessing how variables covary underlies such fundamental behaviors as learning, categorization, and judging causation. For our purposes, what is important about the normative model for this task is that all four cells are equally important (see Chapter 11 for details).

In contrast to the normative model, probably the most robust finding in studies of covariation assessment is that participants do not find the four cells equally important. In particular, Cell A has the largest impact on behavior and Cell D has the smallest impact (Jenkins & Ward, 1965; Kao & Wasserman, 1993; Levin, Wasserman, & Kao, 1993; Lipe, 1990; Smedslund, 1963; Ward & Jenkins, 1965; Wasserman, Dorner, & Kao, 1990). The typical order in terms of impact is $A > B \approx C > D$. This differential impact of the four cells is routinely interpreted as nonnormative. For example, Kao and Wasserman (1993, p. 1365) state that, "It is important to recognize that unequal utilization of cell information implies that nonnormative processes are at work," and Mandel and Lehman (1998) attempted to explain differential cell utilization in terms of a combination of two reasoning biases.

If one views a covariation task as testing the hypothesis that there is a positive relationship between the two variables, then Cells A and D are evidence for the hypothesis and Cells B and C are evidence against it. Note that the larger impact of Cell A compared to Cell D – both of which provide confirming evidence for a positive relationship – is analogous to the larger impact of confirming observations that are mentioned in the hypothesis compared to those that are not mentioned. Also analogous is that one can adopt a Bayesian view of the covariation task in which participants (a) view the four

cell observations as a sample from a larger population of interest and (b) assume that the presence of variables is rare ($p < 0.5$) and their absence common ($p > 0.5$) in the larger population. Note that this latter assumption is related to, but different from, the rarity assumption discussed earlier, which regarded how conditional rules or hypotheses are phrased.

Under these assumptions, Cell A (joint presence) is normatively more informative than Cell D (joint absence) for determining if there is a relationship between the variables rather than no relationship. Observing the rare observation, Cell A, distinguishes better between these two possibilities. Similar to the discussion about rarity earlier, if absence of the two variables is common, then it would not be surprising to see both variables absent, a Cell D observation, even if the variables were independent. In contrast, observing their joint presence would be surprising, *especially* if the variables were independent. Furthermore, under the current assumptions, Cells B and C (evidence against a positive relationship) fall in between Cells A and D in terms of informativeness, which is also consistent with the empirical findings. In short, assuming that presence is rare, a Bayesian account can naturally explain the perceived differences in cell informativeness (Anderson, 1990; Anderson & Sheu, 1995; McKenzie & Mikkelsen, 2000, in press).

Is the presence of an event or a feature usually rarer than its absence? That is, might it be adaptive to assume that presence is rare? The answer will depend on the specific circumstances, but in the majority of cases, the answer appears to be yes. Most things are not red, most things are not mammals, most people do not have a fever, and so on. Here's another way to think about the issue: Imagine two terms, "X" and "not-X" (e.g., red things and non-red things, accountants and non-accountants), where there is no simple, non-negated term for not-X. Which would be the larger category, X or not-X? Not-X appears to be the larger category in the vast majority of cases.

McKenzie and Mikkelsen's (in press) results support this idea. They discovered that a Cell A "bias" became a Cell D "bias" when they used variables that were familiar to the participants and it was clear that absence was rare, providing evidence that the robust Cell A bias demonstrated over the past four decades stems from (a) participants' Bayesian approach to the task, and (b) their default assumption (probably implicit) that presence is rare. A Bayesian approach, combined with an eye toward the structure of the natural environment, can help explain how people evaluate covariation evidence.

Environment summary

Several robust "errors" in hypothesis testing and evaluation can be explained by adopting a qualitatively Bayesian perspective that is influenced by the predictable structure of our natural environment. (For more on the influence of the "real world" on people's behavior, see Chapters 4 and 15.) In particular, participants appear to harbor strong assumptions about event rarity that reflect experiences outside the laboratory. Normative principles combined with considerations of the environment can provide compelling explanations of behavior.

Summary and Implications

Because hypothesis development is a ubiquitous behavior and plays a crucial role in understanding our complex world, early claims about confirmation bias were quite alarming. However, there is now a consensus that many of the early confirmation bias claims were overstated. Neither testing nor evaluation strategies alone appear to necessitate confirmation bias, though certain combinations of the two can (Klayman, 1995).

Laboratory studies can indicate people's general tendencies in testing and evaluating hypotheses. Often these tendencies do not coincide with what is considered optimal for the task of interest. What to conclude about real-world behavior based on these studies is not so straightforward, however. Participants' strategies might reflect what works in the real world, if not the particular laboratory task (e.g., Funder, 1987; Gigerenzer, Todd, & the ABC Research Group, 1999; Hogarth, 1981; McKenzie, 1994, 2003). I have argued that many "biases" can be explained by adopting a Bayesian perspective combined with assumptions about event rarity that appear to reflect our natural environment.

It is probably worth noting that I am *not* claiming that people never make errors. For example, if the selection task instructions make clear that only the four cards showing are of interest, then selecting the P and Q cards is an error (in Funder's, 1987, sense of "error"; see also McKenzie, 2003). Similarly, if the covariation task instructions make clear that participants are to summarize the relationship between the variables for only the observations presented, then giving more weight to joint presence observations is an error (McKenzie, 2003; McKenzie & Mikkelsen, in press). However, when discussing errors in their concept attainment tasks, Bruner et al. (1956, p. 240) wrote, "Little is added by calling them errors. They are dependent variables, these tendencies, whose determinants have yet to be discovered." Indeed. I would only add that, when adaptation to the environment is a crucial part of the explanation of such tendencies, even less is added by calling them errors.

In addition, I am not claiming that people are optimal Bayesians (see, e.g., McKenzie, 1994). Instead, I argue that participants are sensitive to the rarity of data, which is normatively defensible from a Bayesian perspective. People appear to behave in a *qualitatively* Bayesian manner in a variety of tasks that were not intended by the experimenters to be Bayesian tasks. The consistency in findings across these different tasks lends credence to the current viewpoint.

Finally, it will be obvious to most readers that I am hardly the first to argue that taking into account the environment is crucial for understanding behavior (see, e.g., Anderson, 1990; Brunswik, 1956; Gibson, 1979; Gigerenzer et al., 1999; Hammond, 1955; Marr, 1982; Simon, 1955, 1956; Toda, 1962; Tolman & Brunswik, 1935). Nonetheless, this approach is not exploited enough in research on judgment and decision making in general, and on hypothesis development in particular. I hope that I have convinced the reader that much can be gained by the approach. When robust "errors" occur in the laboratory, a fruitful strategy is to (a) ask about the conditions under which such behavior would make sense, (b) see if such conditions describe the

natural environment, and (c) test resulting predictions. The importance of such an analysis lies not so much in its implications for how "good" performance is, but in its ability to provide a *deeper understanding* of behavior (see also Anderson, 1990). Almost 50 years ago, Bruner et al. (1956, p. 240) wrote: "These [strategies], though they may lead to inefficient behavior in particular problem-solving situations, may represent highly efficient strategies when viewed in the broader context of a person's normal life." The recent research described above appears to provide strong support for their hypothesis.

Acknowledgment

Preparation of this chapter was supported by NSF grants SES-0079615 and SES-0242049. For their helpful comments, the author thanks Nigel Harvey, Derek Koehler, Mike Liersch, John Payne, and Shlomi Sher.

References

Allan, L. G. (1993) Human contingency judgments: Rule based or associative? *Psychological Bulletin*, 114, 325–448.

Anderson, J. R. (1990) *The Adaptive Character of Thought*. Hillsdale NJ: Erlbaum.

Anderson, J. R. & Sheu, C-.F. (1995) Causal inferences as perceptual judgments, *Memory and Cognition*, 23, 510–24.

Bassok, M. & Trope, Y. (1984) People's strategies for testing hypotheses about another's personality: Confirmatory or diagnostic? *Social Cognition*, 2, 199–216.

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, 223–41.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956) *A Study of Thinking*. New York: Wiley.

Brunswik, E. (1956) *Perception and the Representative Design of Psychological Experiments* (2nd edn.). Berkeley, CA: University of California Press.

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

Evans, J. St. B. T. (1989) *Bias in Human Reasoning: Causes and Consequences*. Hillsdale, NJ: Erlbaum.

Fischhoff, B. & Beyth-Marom, R. (1983) Hypothesis evaluation from a Bayesian perspective, *Psychological Review*, 90, 239–60.

Funder, D. C. (1987) Errors and mistakes: Evaluating the accuracy of social judgment, *Psychological Bulletin*, 101, 75–90.

Garcia, J. & Koelling, R. A. (1966) Relation of cue to consequence in avoidance learning, *Psychonomic Science*, 4, 123–4.

Gibson, J. J. (1979) *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989) *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge: Cambridge University Press.

Gigerenzer, G, Todd, P. M., & the ABC Research Group (1999) *Simple Heuristics that Make Us Smart*. Oxford: Oxford University Press.

Hammond, K. R. (1955) Probabilistic functioning and the clinical method, *Psychological Review*, 62, 255–62.

Hogarth, R. M. (1981) Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics, *Psychological Bulletin*, 47, 116–31.

Horwich, P. (1982) *Probability and Evidence*. Cambridge: Cambridge University Press.

Howson, C. & Urbach, P. (1989) *Scientific Reasoning: The Bayesian Approach*. La Salle, IL: Open Court.

Jenkins, H. M. & Sainsbury, R. S. (1969) The development of stimulus control through differential reinforcement. In N. J. Mackintosh and W. K. Honig (eds.), *Fundamental Issues in Associative Learning*. Halifax, Nova Scotia, Canada: Dalhousie University Press.

Jenkins, H. M. & Sainsbury, R. S. (1970) Discrimination learning with the distinctive feature on positive or negative trials. In D. Mostofsky (ed.), *Attention: Contemporary Theory and Analysis*. New York: Appleton-Century-Crofts.

Jenkins, H. M. & Ward, W. C. (1965) The judgment of contingency between responses and outcomes, *Psychological Monographs: General and Applied*, 79(1, whole no. 594).

Kao, S.-F. & Wasserman, E. A. (1993) Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1363–86.

Klayman, J. (1995) Varieties of confirmation bias, *Psychology of Learning and Motivation*, 32, 385–418.

Klayman, J. & Ha, Y.-W. (1987) Confirmation, disconfirmation, and information in hypothesis testing, *Psychological Review*, 94, 211–28.

Koehler, J. J. (1993) The influence of prior beliefs on scientific judgments of evidence quality, *Organizational Behavior and Human Decision Processes*, 56, 28–55.

Levin, I. P., Wasserman, E. A., & Kao, S.-F. (1993) Multiple methods for examining biased information use in contingency judgments, *Organizational Behavior and Human Decision Processes*, 55, 228–50.

Levine, M. (1971) Hypothesis theory and nonlearning despite ideas S-R reinforcement contingencies, *Psychological Review*, 78, 130–40.

Lipe, M. G. (1990) A lens-model analysis of covariation research, *Journal of Behavioral Decision Making*, 3, 47–59.

Lord, C. G., Ross, L., & Lepper, M. R. (1979) Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence, *Journal of Personality and Social Psychology*, 11, 2098–109.

Mackie, J. L. (1963) The paradox of confirmation, *British Journal for the Philosophy of Science*, 13, 265–77.

Mackie, J. L. (1974) *The Cement of the Universe: A Study of Causation*. Oxford: Clarendon.

Mandel, D. R. & Lehman, D. R. (1998) Integration of contingency information in judgments of cause, covariation, and probability, *Journal of Experimental Psychology: General*, 127, 269–85.

Marr, D. (1982) *Vision*. San Francisco: W. H. Freeman.

McKenzie, C. R. M. (1994) The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference, *Cognitive Psychology*, 26, 209–39.

McKenzie, C. R. M. (2003) Rational models as theories – not standards – of behavior, *Trends in Cognitive Sciences*, 7, 403–6.

McKenzie, C. R. M., Ferreira, V. S., Mikkelsen, L. A., McDermott, K. J., & Skrable, R. P. (2001) Do conditional hypotheses target rare events? *Organizational Behavior and Human Decision Processes*, 85, 291–309.

McKenzie, C. R. M., Lee, S. M., & Chen, K. K. (2002). When negative evidence increases confidence: Change in belief after hearing two sides of a dispute, *Journal of Behavioral Decision Making, 15*, 1–18.

McKenzie, C. R. M. & Mikkelsen, L. A. (2000) The psychological side of Hempel's paradox of confirmation, *Psychonomic Bulletin and Review*, 7, 360–66.

McKenzie, C. R. M. & Mikkelsen, L. A. (in press) A Bayesian view of covariation assessment, *Cognitive Psychology*.

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, 85–95.

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.

Newman, J., Wolff, W. T., & Hearst, E. (1980) The feature-positive effect in adult human subjects, *Journal of Experimental Psychology: Human Learning and Memory*, 6, 630–50.

Nickerson, R. S. (1996) Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation, *Thinking and Reasoning*, 2, 1–31.

Nickerson, R. S. (1998) Confirmation bias: A ubiquitous phenomenon in many guises, *Review of General Psychology*, 2, 175–220.

Nisbett, R. & Ross, L (1980) *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Oaksford, M. & Chater, N. (1994) A rational analysis of the selection task as optimal data selection, *Psychological Review*, 101, 608–31.

Oaksford, M. & Chater, N. (2003) Optimal data selection: Revision, review, and re-evaluation, *Psychonomic Bulletin and Review*, 10, 289–318.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993) *The Adaptive Decision Maker*. Cambridge: Cambridge University Press.

Poletiek, F. (2001) *Hypothesis-testing Behavior*. East Sussex: Psychology Press.

Popper, K. R. (1959) *The Logic of Scientific Discovery*. New York: Harper & Row.

Restle, F. (1965) Significance of all-or-none learning, *Psychological Bulletin*, 64, 313–25.

Simon, H. A., (1955) A behavioral model of rational choice, *Quarterly Journal of Economics*, 69, 99–118.

Simon, H. A. (1956) Rational choice and the structure of the environment, *Psychological Review*, 63, 129–38.

Skov, R. B. & Sherman, S. J. (1986) Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation, *Journal of Experimental Social Psychology*, 22, 93–121.

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992) Information selection and use in hypothesis testing: What is a good question and what is a good answer? *Memory and Cognition*, 20, 392–405.

Smedslund, J. (1963) The concept of correlation in adults, *Scandinavian Journal of Psychology*, 4, 165–73.

Sniezek, J. A. & Naylor, J. C. (1978) Cue measurement scale and functional hypothesis testing in cue probability learning, *Organizational Behavior and Human Performance*, 22, 366–74.

Snyder, M. (1981) Seek and ye shall find: Testing hypotheses about other people. In E. T. Higgins, C. P. Heiman, and M. P. Zanna (eds.), *Social Cognition: The Ontario Symposium on Personality and Social Psychology* (pp. 277–303). Hillsdale, NJ: Erlbaum.

Snyder, M. & Campbell, B. H. (1980) Testing hypotheses about other people: The role of the hypothesis, *Personality and Social Psychology Bulletin*, 6, 421–6.

Snyder, M. & Swann, W. B., Jr. (1978) Hypothesis-testing in social interaction, *Journal of Personality and Social Psychology*, 36, 1202–12.

Thagard, P. (1989) Explanatory coherence, *Behavioral and Brain Sciences*, 12, 435–502.

Toda, M. (1962) The design of a fungus-eater: A model of human behavior in an unsophisticated environment, *Behavioral Science*, 7, 164–83.

Tolman, E. C. & Brunswik, E. (1935) The organism and the causal structure of the environment, *Psychological Review*, 42, 43–77.

Trope, Y. & Bassok, M. (1982) Confirmatory and diagnosing strategies in social information gathering, *Journal of Personality and Social Psychology*, 43, 22–34.

Trope, Y. & Bassok, M. (1983) Information-gathering strategies in hypothesis testing, *Journal of Experimental Social Psychology*, 19, 560–76.

Ward, W. C. & Jenkins, H. M. (1965) The display of information and the judgment of contingency, *Canadian Journal of Psychology*, 19, 231–41.

Wason, P. C. (1960) On the failure to eliminate hypotheses in a conceptual task, *Quarterly Journal of Experimental Psychology*, 12, 129–40.

Wason, P. C. (1966) Reasoning. In B. M. Foss (ed.), *New Horizons in Psychology* (pp. 135–51). Harmondsworth, England: Penguin.

Wason, P. C. (1968) Reasoning about a rule, *Quarterly Journal of Experimental Psychology*, 20, 273–81.

Wasserman, E. A., Dorner, W. W., & Kao, S.-F. (1990) Contributions of specific cell information to judgments of interevent contingency, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 509–21.

Zuckerman, M., Knee, C. R., Hodgins, H. S., & Miyake, K. (1995) Hypothesis confirmation: The joint effect of positive test strategy and acquiescence response set, *Journal of Personality and Social Psychology*, 68, 52–60.