COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 35 (2011) 499–526

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2010.01161.x

Seeking Confirmation Is Rational for Deterministic Hypotheses

Joseph L. Austerweil, Thomas L. Griffiths

Department of Psychology, University of California, Berkeley

Received 31 October 2008; received in revised form 25 June 2010; accepted 15 October 2010

Abstract

The tendency to test outcomes that are predicted by our current theory (the confirmation bias) is one of the best-known biases of human decision making. We prove that the confirmation bias is an optimal strategy for testing hypotheses when those hypotheses are deterministic, each making a single prediction about the next event in a sequence. Our proof applies for two normative standards commonly used for evaluating hypothesis testing: maximizing expected information gain and maximizing the probability of falsifying the current hypothesis. This analysis rests on two assumptions: (a) that people predict the next event in a sequence in a way that is consistent with Bayesian inference; and (b) when testing hypotheses, people test the hypothesis to which they assign highest posterior probability. We present four behavioral experiments that support these assumptions, showing that a simple Bayesian model can capture people's predictions about numerical sequences (Experiments 1 and 2), and that we can alter the hypotheses that people choose to test by manipulating the prior probability of those hypotheses (Experiments 3 and 4).

Keywords: Confirmation bias; Rational analysis; Hypothesis testing; Decision making; Bayesian inference; Determinism

1. Introduction

How *should* a scientist seek evidence to help her find the hypothesis that explains a phenomenon? Over the last century, philosophers of science have explored this question. For example, Popper (1935/1990) argued that scientists ought to follow the strategy of *falsification*, seeking evidence most likely to invalidate their current theory. Popper's main argument examines the fundamental difference between two types of evidence a scientist

Correspondence should be sent to Joseph Austerweil, Department of Psychology, University of California, Berkeley, 3210 Tolman Hall #1650, Berkeley, CA 94720-1650. E-mail: joseph.austerweil@gmail.com

can receive from an experiment: verifying and falsifying. A scientist discovers *verifying* evidence from an experiment when the result of the experiment is predicted by her hypothesis. Though the verifying evidence supports her hypothesis, it will not rule out other alternative hypotheses that also predict the evidence. On the other hand, she discovers *falsifying* evidence from her experiment if its result is not predicted by her hypothesis. As falsifying evidence invalidates her hypothesis and verifying evidence only supports it (while not ruling out alternative accounts), she should seek this more powerful falsifying evidence when testing her hypothesis.

For example, imagine that a scientist was investigating the link between blood pressure and heart attacks (before it was discovered that such a relationship exists). After observing a patient with a blood pressure of 180 mmHg have a heart attack, the scientist forms the hypothesis that an average blood pressure of 180 mmHg or higher can lead to a heart attack (and that having less than an average blood pressure of 180 mmHg has no relationship with heart attacks). To test this hypothesis, she counts how many people with average blood pressures of 180 mmHg or higher have heart attacks. By noticing that heart attacks are common among people who have an average blood pressure of 180 mmHg or higher, she verifies her hypothesis. Despite a large amount of evidence that is true under her hypothesis, this seems to be an unwarranted conclusion. If she had sought evidence that is not true under her hypothesis (e.g., looking at people who have an average blood pressure of 175 mmHg), she would have falsified it and could go on to formulate a correct hypothesis about the causes of heart attacks. Thus, it would seem that scientists ought to test their hypotheses by trying to falsify them.

This type of argument about the kind of evidence that people should seek has been extremely influential in the psychological decision-making literature. Interested in whether people adhere to the strategy of falsification, Wason (1960) investigated how people intuitively test their theories about the world. In Wason's classic 2-4-6 task, participants were asked to uncover a rule relating sequences of three numbers after being told that one triplet, (2,4,6), conforms to the rule. The true rule, increasing numbers, subsumes most potential rules (e.g., two more than the previous number) with every triplet predicted by these rules also being valid under the increasing numbers rule. Thus, the true rule can only be discovered by testing numbers that are not predicted by one's current best guess at the rule. This falsification-oriented strategy is known in the literature as the *negative test strategy* or NTS (Klayman & Ha, 1987). Rather than follow the NTS, participants choose to test triplets predicted by their current hypothesis (known as the *positive test strategy* or PTS) even though it is impossible to find the true rule this way. For example, many participants in the 2-4-6 task followed the PTS by entertaining the hypothesis that each number is two more than the previous number and testing sequences consistent with this hypothesis, such as (1,3,5). The tendency to follow the PTS is just one instance of what has become known as the *confirmation bias*: the general human tendency to interpret and seek evidence fitting their current theory differently from evidence against it (Klayman & Ha, 1987).

In this paper, we outline a set of environmental conditions under which the PTS is actually an optimal strategy, providing a way to understand why people might pursue this strategy. Previous work has identified settings in which following either the PTS or NTS is

more likely to yield falsification (e.g., Klayman & Ha, 1987). However, this normative analysis produces predictions that are quite different from human behavior. For example, people still use positive tests in situations where negative tests are more likely to yield falsification, such as those encountered in Wason's (1960) experiment. We complement this analysis by showing that the PTS is more likely to yield falsification and optimally reduces uncertainty provided the world is inherently *deterministic* (i.e., given the rule is true, there is only one possible next outcome). This suggests we could explain the use of the PTS as the result of an assumption of determinism on the part of human learners, consistent with recent results showing that children assume that many causal relationships are deterministic (e.g., Gelman, Coley, & Gottfried, 1994; Schulz & Sommerville, 2006). This emphasis on the structure of the environment parallels similar strategies pursued in other rational analyses of how people seek information (e.g., Oaksford & Chater, 1994).

The plan of the paper is as follows. First, we introduce the task of predicting the next event in a sequence. Under the assumption that hypotheses are deterministic (given a sequence of events, a hypothesis predicts only one next event), we prove that the PTS is optimal in many situations. Next, we define a Bayesian model of sequence prediction for numerical stimuli and use two behavioral experiments to show that it captures human predictions. If people are seeking evidence optimally, then they should choose to verify the next number predicted by the hypothesis they believe is most likely. Thus, our theoretical analyses predict that the subjective probability of hypotheses should affect the way people seek evidence in addition to their sequential predictions. In two other experiments, we demonstrate that changing a person's beliefs about the probability of hypotheses affects the evidence they seek. We conclude by discussing the implications of our results and how they relate to previous work.

2. Analyzing confirmation with deterministic hypotheses

Wason's (1960) original demonstration of the confirmation bias used a task in which people tried to identify an abstract rule relating sets of three numbers. However, a bias toward confirmation can appear in any situation in which people need to test hypotheses. In this paper, we consider the effects of different strategies in the related task of *sequence prediction*. Given a sequence of events, how do we predict what will occur next? For example, suppose you see a woman outside an airport and then at the security checkpoint. How likely is it that she stays at the security checkpoint (she is a security guard), walks to a gate waiting area (she is a passenger), or to a gate ticket collection booth (she is a crewmember)? We use the sequence prediction task because it allows us to explore the consequences of making different assumptions about the world in a way that is more natural than Wason's original task. In particular, it is possible that the hypotheses under consideration are deterministic, predicting exactly one outcome. In our example, it might be appropriate to assume that if the woman is a crewmember, the probability she walks to the gate ticket collection booth is one, and likewise for all other hypotheses to their respective locations. Clearly, the probability of each possible next event depends on the

probability of the hypotheses explaining the observed events and the probability of the next event under these hypotheses. As there is no means of predicting the next event with complete certainty, this is an inductive problem.

This problem can be expressed in terms of probability theory. Given a sequence of previous events or objects ($\vec{x} = (x_1, \ldots, x_{i-1})$) the probability of a next event (x_i) is

$$P(x_i \mid \vec{x}) = \sum_h P(x_i \mid h, \vec{x}) P(h \mid \vec{x}), \tag{1}$$

where $P(x_i \mid h, \vec{x})$ is the probability of the next event under hypothesis h, and $P(h \mid \vec{x})$ is the posterior probability of that hypothesis given the sequence $\vec{x}$. This posterior probability can be obtained from Bayes' rule, with

$$P(h \mid \vec{x}) = \frac{P(\vec{x} \mid h) P(h)}{\sum_{h'} P(\vec{x} \mid h') P(h')} \tag{2}$$

being the normalized product of the likelihood, $P(\vec{x} \mid h)$, and the prior probability of the hypothesis $P(h)$. For the above example, the probability that the woman is a security guard instead of a passenger depends on the relative probabilities of a security guard and a passenger going to the security checkpoint and the base rates with which passengers and security guards appear at the airport.

Suppose we now meet the woman's husband, and we get to ask him one (yes or no) question about where she will be next. What is the best question to ask in order to discover her role (i.e., whether she's a security guard, passenger, or crewmember)? This is equivalent to a scientist determining the best question to test her hypothesis. In the remainder of this section, we show that there is a simple answer to this question provided our hypotheses are deterministic, allowing only one value for x given $\vec{x}$. In this case, the positive test strategy (asking about the event that corresponds to the most probable hypothesis) is optimal. Thus, the best question to ask is our best guess about where the woman will be next.

We will analyze two methods for identifying which question we should ask. The first is based on falsification—picking the question that is most likely to yield falsifying evidence (Popper, 1935/1990). With both strategies, we can falsify our hypothesis, thus we should pick the question that falsifies our hypothesis with highest probability (Klayman & Ha, 1987). If we believe the woman is a security guard, which question should we ask her husband to yield falsifying evidence? In other words, should we ask her husband if she will be at the security checkpoint, the gate sitting area, or the ticket collection counter? If we ask her husband whether or not she is at the security checkpoint (PTS), we falsify our hypothesis when her husband responds that she is not there (meaning she is somewhere at a gate). If we ask her husband whether or not she is at the gate waiting area or the gate ticket collection counter (NTS), we falsify our hypothesis when she is there. If maximizing the probability of falsification is our goal, then intuitively (although we will prove it in the subsequent sections) we should use the PTS because the probability that PTS yields falsification is equal to the sum of the probabilities of falsifying with all the different possible NTS questions.

The second method of identifying which question to ask that we analyze is a measure based on information theory, the *expected information gain* (EIG; Klayman, 1987; Oaksford & Chater, 1994). According to information theory (Shannon, 1948), the *entropy*

$$H(P(x)) = -\sum_x P(x) \log_2 P(x)$$

measures the amount of randomness in a probability distribution *P(x)*. For example, the entropy of a fair coin is 1 ($-.5 \log_2 .5 + .5 \log_2 .5 = 1$) and the entropy of a two-headed coin is 0 ($-1 \log_2 1 + 0 \log_2 0 = 0$, where $0 \log_2 0$ is defined to be 0). This matches our intuition that we are far more certain of the outcome from the toss of a two-headed coin. The amount of information gained from observing an outcome is the difference between the entropy of the distribution characterizing our beliefs before and after that observation. Thus, the information gained about the a set of hypotheses for which our current beliefs are described by the posterior distribution $P(h \mid \vec{x})$, given a sequence of objects from performing a test c and learning its outcome r, is

$$I(P(h \mid \vec{x}), P(h \mid \vec{x}, r, c)) = H(P(h \mid \vec{x})) - H(P(h \mid \vec{x}, r, c)),$$

where $P(h \mid \vec{x}, r, c)$ reflects the information provided by (r,c),

$$P(h \mid \vec{x}, r, c) = \frac{P(r \mid h, c, \vec{x}) P(h \mid \vec{x})}{P(r \mid c, \vec{x})}$$

with

$$P(r \mid c, \vec{x}) = \sum_h P(r \mid h, c, \vec{x}) P(h \mid \vec{x})$$

being the probability of the outcome r from the test c given our previous observations $\vec{x}$. In sequence prediction, the outcome of a test is either that the queried event is next in the sequence or not. The probability of a positive response ($r = +$) to a query c is simply the probability that c is the next event in the sequence, which depends on h and $\vec{x}$.

As the outcome of a test is unknown prior to performing the test, the information gain cannot be used directly. Instead, we define the optimal test to be the test that has the largest expected information gain. The optimal choice $\hat{c}$ is

$$\hat{c} = \arg\max_c E_{r|c,\vec{x}}[I(P(h \mid \vec{x}), P(h \mid \vec{x}, r, c))],$$

where $E_r[f(r)] = \sum_r f(r) P(r)$ is the expectation of the function f with respect to the distribution P. This reduces to

$$\begin{aligned} \hat{c} &= \arg\max_c \sum_r [H(P(h \mid \vec{x})) - H(P(h \mid \vec{x}, r, c))] P(r \mid c, \vec{x}) \\ &= \arg\min_c \sum_r H(P(h \mid r, c, \vec{x})) P(r \mid c, \vec{x}) \end{aligned}$$

being that choice which minimizes uncertainty after the response.

Maximizing expected information gain has recently been used to explain how people gather information about their hypotheses. Wason (1968b) famously demonstrated that participants do not always conform to the laws of propositional logic when testing a rule, using a card selection task. In this task, participants were asked to test a rule of the form $P \rightarrow Q$, and they were shown four cards that provided information about P on one side and Q on the other. Participants were likely to check the card indicating P was true to see whether the other side indicated that Q was true, but they rarely checked the card indicating Q was false in order to determine whether the other side indicated that P was true (which would violate the rule). Instead, they checked the card showing that Q was true. Rather than focusing on propositional logic as a normative standard for the task, Oaksford and Chater (1994) demonstrated that the tendency to test P and Q shown by participants is normatively prescribed by EIG. Additionally, EIG predicted that testing strategies should be sensitive to the prior probabilities of P and Q, which was demonstrated empirically (Oaksford, Chater, Grainger, & Larkin, 1997). Nelson (2005) found that EIG performs well as a predictor of people's judgments of the usefulness of questions, although other work suggests that people might not always use EIG when testing the appropriateness of rules (Nelson & Movellan, 2001).

Instead of proving directly that the PTS is optimal given that the only possible rules are deterministic, we will proceed in two stages. As the more general proof is conceptually the same (but technically more involved), we first show how the result holds when the hypotheses are deterministic and mutually exclusive. Afterwards, we generalize this constrained proof to cover situations in which the possible hypotheses are any set of deterministic hypotheses.

2.1. The special case of hypotheses mutually exclusive on the next event

By assumption, our hypothesis space contains only deterministic hypotheses. In this section, we further constrain these hypotheses to all make mutually exclusive predictions for the next event in the observed sequence. For example, if we were trying to predict the next number that would appear in a sequence, and we had observed $\vec{x} = (3, 5)$, our hypothesis space could include hypotheses corresponding to both the rules ''+2'' and ''sum of the last two numbers'' because they make mutually exclusive predictions for the next number (7 and 8, respectively) but could not include both ''+2'' and ''increasing prime numbers'' because both predict 7. Under these conditions, every test is a positive test for some hypothesis, and a positive response from such a test yields conclusive verification of the tested hypothesis, while a negative response falsifies the tested hypothesis but is ambiguous about all other hypotheses. We show that testing the event predicted by the a posteriori most probable hypothesis maximizes both the probability of falsifying that hypothesis and the expected information gain.

2.2. Maximizing probability of falsification

Using maximizing the probability of falsifying the current working hypothesis as our normative standard (as in Klayman & Ha, 1987), the analysis is straightforward. The

probability that testing the choice c, consistent with the most probable hypothesis h^c (and only h^c), falsifies that hypothesis is $1 - P(h^c \mid \vec{x})$. The probability of falsifying the hypothesis h^c by testing the choice a predicted by some alternate hypothesis h^a is $P(h^a \mid \vec{x})$. As $1 - P(h^c \mid \vec{x})$ is the sum of the posterior probabilities of all alternate hypotheses, the probability of falsifying a hypothesis by testing the choice predicted by some alternate hypothesis, $P(h^a \mid \vec{x})$, is one component of that sum (i.e., $1 - P(h^c \mid \vec{x}) = \sum_{a \neq c} P(h^a \mid \vec{x}) \geq P(h^a \mid \vec{x})$). Thus, to maximize the probability of falsifying the working hypothesis, you should test the choice predicted by the working hypothesis or use the PTS.

2.3. Maximizing the expected information gain

The same result holds when we take maximizing the expected information gain as our goal. As shown above, maximizing the EIG is equivalent to minimizing the expected entropy of the posterior distribution informed by the results of the test. As the hypotheses all predict different next events, if we learn that c is in fact the next event, than we know with certainty that its corresponding hypothesis is true, resulting in an entropy of 0. Thus, the expected entropy reduces to the product of the posterior probability that the tested hypothesis is false and the entropy of the renormalized posterior without the tested hypothesis

$$H\left(\frac{P(h \mid \vec{x})}{1 - P(h^c \mid \vec{x})}\right)(1 - P(h^c \mid \vec{x}))$$

where h^c is the hypothesis corresponding to the choice C. This simplifies to

$$\begin{aligned} &-(1 - P(h^c \mid \vec{x})) \sum_{a \neq c} \tfrac{P(h \mid \vec{x})}{1 - P(h^a \mid \vec{x})} \log_2 \tfrac{P(h \mid \vec{x})}{1 - P(h^a \mid \vec{x})} \\ &\quad = -\sum_{a \neq c} P(h \mid \vec{x}) \log_2 P(h \mid \vec{x}) + \sum_{a \neq c} P(h \mid \vec{x}) \log_2 (1 - P(h^a \mid \vec{x})) \end{aligned}$$

The first of the two sums is the entropy of the posterior without the contribution from the tested hypothesis, and the second simplifies because the log portion does not vary over the sum. Consequently, we can rewrite this quantity as

$$H(P(h \mid \vec{x})) + P(h^c \mid \vec{x}) \log_2 P(h^c \mid \vec{x}) + (1 - P(h^c \mid \vec{x})) \log_2 (1 - P(h^c \mid \vec{x}))$$

As the entropy of the posterior does not depend on the choice c, it does not influence the optimal choice. This means that the choice that maximizes the EIG is

$$\hat{c} = \arg\min_c P(h^c \mid \vec{x}) \log_2 P(h^c \mid \vec{x}) + (1 - P(h^c \mid \vec{x})) \log_2 (1 - P(h^c \mid \vec{x}))$$

which is the negative entropy of a distribution in which h^c and its alternatives are the only two possible outcomes.

The entropy of a distribution is concave (there is one global maximum) and is maximized when the distribution is uniform (Cover & Thomas, 1991). Thus, the optimal strategy is to

make the choice corresponding to the hypothesis with posterior probability closest to .5. It is easy to show that this is the hypothesis with highest posterior probability.[1] There are two cases. If all probabilities $P(h \mid \vec{x})$ are less than .5, then the hypothesis for which $P(h \mid \vec{x})$ is greatest is clearly the closest to .5. If the probability of some hypothesis is greater than .5, there is only one such hypothesis, and the distance of the probability of all other hypotheses from .5 will be at least as great, as these hypotheses divide the remaining probability mass. Thus, confirmation—choosing to test the hypothesis with highest posterior probability—maximizes the EIG.

2.4. Generalization to any set of deterministic hypotheses

We now generalize this analysis for both maximizing the probability of falsification and the EIG by relaxing the assumption that all hypotheses must predict different next numbers. In other words, if $\vec{x} = (3, 5)$, our hypothesis space can be ''+2'' and ''increasing prime numbers,'' or any countably infinite space of deterministic hypotheses. In this more general case, the PTS is still optimal for maximizing the probability of falsification; however, there are cases where the PTS may be suboptimal for EIG.

2.5. Maximizing the probability of falsification

Similar to the previous analysis, we demonstrate that the probability of falsifying the working hypothesis with the NTS is one component of the total probability of falsifying with the PTS and thus the PTS is optimal. In general, let $\mathscr{H}^c$ be the set of hypotheses that predict c as the next event. The most probable hypothesis is a member of this set, but unlike the constrained analysis other hypotheses can be in $\mathscr{H}^c$. The probability that the PTS yields falsification is the posterior probability that c is not the next event in the sequence or $1 - P(\mathscr{H}^c \mid \vec{x}) = \sum_{a \neq c} P(\mathscr{H}^a \mid \vec{x})$. The probability that the NTS yields falsification is the posterior probability that the alternate event a (not predicted by the working hypothesis) is the next event in the sequence or $P(\mathscr{H}^a \mid \vec{x})$. As the probability of falsifying with NTS is the posterior probability of only one alternate event and the probability of PTS includes the posterior probability of all alternate events (not the working hypothesis), the PTS is always as good or better than the NTS.

2.6. Maximizing the expected information gain

We can now generalize this analysis for the EIG, relaxing the assumption that all hypotheses make distinct predictions for the next event. In the general case, every choice c partitions the hypothesis space into two sets. Let $\mathscr{H}^c$ be the set of hypotheses that predict c as the next event and $\mathscr{H}^{\bar{c}}$ be the set of hypotheses that do not. The set that makes the wrong prediction will be eliminated, receiving probability 0, and the set that makes the right prediction will have their posterior probabilities renormalized. The analysis then proceeds similarly to the derivation given above, replacing h^c with $\mathscr{H}^c$, with $P(\mathscr{H}^c|\vec{x}) = \sum_{h \in \mathscr{H}^c} P(h \mid \vec{x})$, although there is an extra wrinkle produced by the fact that

confirmation does not guarantee an entropy of 0. Analogously, choosing the choice that maximizes the EIG is equivalent to $\arg\max_c H([P(\mathscr{H}^c \mid \vec{x}), 1 - P(\mathscr{H}^c \mid \vec{x})])$. Thus, the optimal test is that which produces $P(\mathscr{H}^c \mid \vec{x})$ closest to .5. If there is a single hypothesis with posterior probability greater than or equal to a half, then confirming that hypothesis (which is the current best hypothesis) is the optimal strategy. If this is not the case, confirming the current best hypothesis can be suboptimal, as it may be possible to construct an amalgam of hypotheses that agree on some c and have posterior probabilities that sum to a value closer to .5. However, such circumstances are unusual, and our result thus indicates that in many cases where we believe there is a rule governing a sequence of events, the PTS is also optimal according to maximizing EIG.

2.7. Summary

According to the two normative standards, when discovering what rule underlies the events you observe, testing the event predicted by your most probable hypothesis is rational—as long as the rules are deterministic. This result is similar to that obtained by Oaksford and Chater (1994) in their analysis of Wason's (1968b) card selection task, which we introduced earlier in the paper. By assuming that the consequent and antecendent of rules are rare and that participants are trying to become as certain as possible of whether or not a rule applies (i.e., that they are maximizing EIG), Oaksford and Chater (1994) provided a rational justification for behavior that seems irrational from a logical perspective. The results of Klayman and Ha (1987) also have a similar character, demonstrating that the PTS is rational (in terms of maximizing the probability of falsification) when the objects that rules act on in the world are rare. Our analysis replaces the rarity of predicates with deterministic rules, providing another way to understand why people might pursue a strategy of confirmation (the PTS) rather than falsification (the NTS).

In order for this formal analysis of the consequences of determinism to provide a potential explanation for why people choose to confirm hypotheses rather than falsify them, we need to show that the assumptions that it makes about human inferences are plausible. In particular, we need to justify two assumptions: that people predict the next event in a sequence in a way that can be described in terms of Bayesian inference, and that they test the rule with highest posterior probability. In the remainder of the paper, we develop a simple Bayesian model for predicting numerical sequences, and we use a series of experiments to examine the adequacy of this model as a characterization of human judgments and the extent to which people are sensitive to the posterior probability of rules in selecting their tests.

Importantly, we are not proposing that our Bayesian model is an account of how people actually predict the next event in a sequence of events. The model is intended as a computational-level analysis in the sense introduced by Marr (1982): an analysis of how an ideal solution to the problem of sequence prediction should look. Appealing to Marr's classic analogy, it is useful to understanding human sequence prediction in the same sense that the physics of aerodynamics is useful for understanding bird flight. Our analysis provides a way to understand the confirmation bias as a rational strategy for testing hypotheses in a

deterministic sequential world. We are not claiming it is an accurate description of human sequence prediction at the process or algorithmic level.[2] Exploring the mechanisms and processes by which people predict sequential events and test hypotheses is an interesting avenue for future work, and we consider some possibilities in Section 8.

3. A Bayesian model for predicting numerical sequences

The analysis of the positive test strategy outlined above relies upon the assumption that we can accurately characterize people's predictions about sequences in terms of Bayesian inference. In the remainder of the paper, we develop a Bayesian model of a particular kind of sequence prediction—prediction of the next element in a sequence of numbers—and use this model to test this basic assumption, and to show that people are sensitive to the relative probabilities of different hypotheses in exactly the way that this account predicts. The sole purpose of developing this Bayesian model and conducting these experiments is to test the predictions of our analysis. To do this, we first define a simple Bayesian model for predicting numerical sequences, postulating a hypothesis space that might capture people's judgments in this task. We then estimate appropriate prior probabilities for these hypotheses from human judgments in Experiments 1 and 2. Experiments 3 and 4 manipulate those prior probabilities and demonstrate that people adapt their tests to their environment as our analysis predicts.

The numerical sequence prediction task we use is inspired by Wason's (1960) 2-4-6 task, in that the domain is ordered sequences of numbers. However, the potential space of hypotheses in this case is more constrained, with relevant rules being easier to identify. This makes it easier to test the predictions produced by our analysis. The rules in this domain (e.g., two more than the last number and increasing numbers) include some of the rules that participants found in Wason's experiment (Wason, 1968a). However, they do not include other rules that participants in Wason's experiment found (e.g., the middle number is the average of the outer two numbers). For example, consider the sequence (2,4). Some rules predict the same next number (e.g., $\times 1 + 2$ and *sum of the last two numbers*) and others a different next number (e.g., increasing powers of two). Given a set of hypotheses defined on number sequences, Eqs. 1 and 2 provide the hypothesis with highest posterior probability and the most likely next number, respectively. For the example of the sequence (2,4), these equations would assign highest posterior probability to whichever rule had highest prior probability, and the probability of 6 being the next number would be the sum of the posterior probabilities of $\times 1 + 2$ and *sum of the last two numbers*, while the probability of 8 being the next number would be the posterior probability of increasing powers of two.

Our model assumes that the sequence of observed numbers, $\vec{x} = (x_1, \ldots, x_{i-1})$, is generated from some relational rule h, and that people try to identify this rule based on the observed sequence in order to make accurate predictions. The model is based upon the concept learning framework presented in Tenenbaum (1999) and Tenenbaum and Griffiths (2001), a version of which was applied to a simple ''number game'' similar to our task. In this model, a hypothesis or concept is a set of numbers. Given a set of observed numbers,

the probability that another number is in the set is the sum of the posterior probability of all hypotheses given the observed numbers. Although this model captures people's generalization judgments (e.g., given 8 is in the set, what is the probability that 16 is in the set?), it does not allow for inferences about sequences of numbers. Thus, we extend this Bayesian model to make predictions about sequences. The goal of the model is not to capture all the intricacies of human sequence prediction, but rather to show that people approximate Bayesian inference, and to estimate prior probabilities for a set of hypotheses that allow us to compare the predictions of our theoretical analysis to human hypothesis testing.

Instead of defining the hypotheses as sets of numbers, each hypothesis h is a rule linking k_h previous numbers to the possible next numbers of the sequence. Each hypothesis is associated with a probability distribution over the next number given the previous k_h observed numbers, defining the likelihood function that will be used with that hypothesis in Bayesian inference. For example, ''add three to previous number'' would be associated with a distribution that is one for $x_i = x_{i-1} + 3$ and zero otherwise. Based on these distributions, we divide the types of hypotheses into two separate categories: deterministic and nondeterministic. A *deterministic* hypothesis, such as ''add three to previous number,'' has only one correct next number and conforms to the following form: $h(x_{i-1}, \ldots, x_{i-k+1})$: $\mathscr{Z}^k \rightarrow \mathscr{Z}$, where $\mathscr{Z}$ is the set of integers. In other words, a deterministic hypothesis is a function from sets of k_h numbers to a single number. For example, the likelihood function for the ''sum of the last two numbers'' rule (Fibonacci sequence, $k_h = 2$) is:

$$P(x_i|h, x_{i-1}, x_{i-2}) = \begin{cases} 1 & \text{if } x_i = x_{i-1} + x_{i-2}, \\ 0 & \text{otherwise} \end{cases}$$

Conversely, a *nondeterministic* hypothesis allows the next number to take more than one value. For example, the following likelihood function models the ''increasing numbers'' rule ($k_h = 1$):

$$P(x_i \mid h, x_{i-1}; v) = \begin{cases} \frac{1}{v+1} & x_i \geq x_{i-1} \wedge x_i - x_{i-1} \leq v, \\ 0 & \text{otherwise,} \end{cases}$$

where v is the largest increase possible from the last number (v was fixed to 65 for all experiments). As all the rules are based on preceding numbers, we also need a scheme for generating the initial numbers in a sequence. We do this by sampling the first number in the sequence, x_0, from a distribution assigning probability $1/|1 + x_0|$ to the entire set of positive and negative integers. For rules that require multiple initial numbers to generate the next number (i.e., with $k_h > 1$), we sampled the next number in the sequence from the same distribution, but now centered it around the last number.

To develop a model that describes human judgments, we need to define a set of hypotheses that capture the kinds of regularities that people expect to see expressed in sequences. Our goal is not to be exhaustive in listing all of the hypotheses that people consider, but to cover the possibilities with sufficient resolution to allow us to test the predictions our formal account makes about the confirmation bias. The hypotheses we used are partitioned

into seven different sets of the same rule type: ''$\times C + K$,'' ''sum of the last two numbers,'' pairwise mixtures of ''$\times C + K$'' rules (with the next number chosen with equal probability from one of two rules each time), ''repeat the last k_h numbers,'' ''the kth power'' (for $k = 2$ and 3), ''consecutive prime numbers'' (starting at three), and the random rules (''decreasing,'' ''increasing,'' and ''random numbers''). The ''$\times C + K$'' hypotheses cover any rule of the form $x_i = Cx_{i-1} + K$ (e.g., ''$\times 1 + 3$'' or ''+3,'' or $\times 1 + 0$, which is ''repeat the same number''), and we considered $C \in \{-3,\ldots,3\}$ except zero, $K \in \{-5, \ldots, 5\}$. In total, this yields 135 hypotheses. The prior probability of all rules of a given type is uniform within that set, and the prior probabilities of the rules of different types are free parameters of the model. As the prior probability distribution defined through the parameters must sum to one, the model thus has six free parameters, which are used to bring its predictions into line with human performance.

The model defined in this section provides all we need to compute the posterior distribution over hypotheses given a sequence of numbers (Eq. 2) and consequently to predict the next number in a sequence (Eq. 1). To generate the posterior distribution over hypotheses, the likelihood of each hypothesis is found by stepping through each number in the sequence. First, starting at the beginning of the sequence, the first numbers (k_h of them) are generated from the initial number distribution until the hypothesis can be applied (e.g., zero numbers for the ''prime number'' hypothesis, one number for the $\times C + K$ rule, two numbers for the ''sum of the last two numbers'' hypothesis).[3] Once the hypothesis can be applied, the probability of each number in the sequence given the previous k_h numbers is multiplied together until the last number of the sequence is included. Finally, the result is multiplied by the prior probability, and then normalized over all hypotheses included in the model.

Before looking at whether or not the hypotheses that people select are sensitive to the manipulation of people's prior beliefs, we need to ensure that the hypotheses included in our model capture human judgments. Thus, Experiments 1 and 2 examine how well this model can characterize the predictions that people make about sequences of numbers. We fit the six parameters defining the prior distribution over hypotheses through comparison to human judgments in Experiment 1. The generalization performance of the model was then assessed by comparing the predictions produced using these fitted values to human performance in Experiment 2, which used a slightly different task and a different set of participants. Experiment 2 thus provided a way to check that the model was not overfitting human responses, as well as a further test of whether people were performing in a way that was consistent with Bayesian inference.

4. Experiment 1: Predicting freeform responses

The analysis presented above demonstrated that the optimal test strategy for testing deterministic rules over sequential events is the PTS. However, our analysis assumes that people can identify the most probable rule and next event given a sequence of events. As this is not a trivial task, our first experiment investigates how participants predict the next number for a given sequence. In particular, we are interested in whether or not these responses can be

captured by a Bayesian model. If so, our analysis is not only a philosophical endeavor but also a justification for human hypothesis-testing behavior. Establishing the prior probabilities of different hypotheses is also a necessary step toward being able to examine how modification of these prior probabilities changes the hypotheses that people choose to test.

In this experiment, participants were asked to predict the next number for five sequences, each generated by a different rule. There were five patterns, four deterministic and one stochastic, each expressed in four sequences of increasing size (length ranging from three to six). The four deterministic patterns were chosen to illustrate how people and the model made judgments for simple and complex rules and when the given sequence was ambiguous as to the underlying rule. The stochastic pattern was chosen to demonstrate that both people and the model make sensible related judgments when the generating rule is not deterministic.

4.1. Methods

4.1.1. Participants

A total of 64 undergraduates from Brown University participated in the experiment for a free ice cream voucher and 82 undergraduates from the University of California, Berkeley participated in the experiment for course credit.

4.1.2. Stimuli

Five relational rules were tested: ''repeat the last number'' or $\times 1 + 0$ (1,1,1,1,1,1—simple), ''sum of the last two numbers'' (1,1,2,3,5,8—complex), ''increasing odd numbers'' (3,5,7,9,11,13—ambiguous), ''increasing prime numbers'' (3,5,7,11,13,17—ambiguous), and ''increasing numbers'' (2,5,17,33,94,100—stochastic).

4.1.3. Procedure

The experiment was conducted as a survey. The four subsequences of each rule were randomly distributed across four different surveys, with each survey containing one subsequence of each rule. Each participant received one survey, with approximately 11 participants seeing each survey. To provide the strongest test of our model, we asked participants to write down what they believed the next number would be, without imposing any constraints on this choice. Participants were told that the sequences may have been generated by a simple relational rule which may not be deterministic, with ''decreasing numbers'' being given as an example, and asked to make predictions for each sequence independently.

4.2. Results and discussion

As shown in Fig. 1, the model and human prediction distributions are in close qualitative correspondence. For example, the model captures extremely well the bimodal nature of the human predictive distribution for the sequence (1,1,2). The predictions shown for the model were obtained by optimizing the prior probability of the different hypothesis types to fit the human data. The estimated prior probabilities of the seven types of hypotheses were .472 for ''$\times C + K$'' as .011 for ''sum of the last two numbers,'' .014 for mixtures of ''$\times C + K$,''

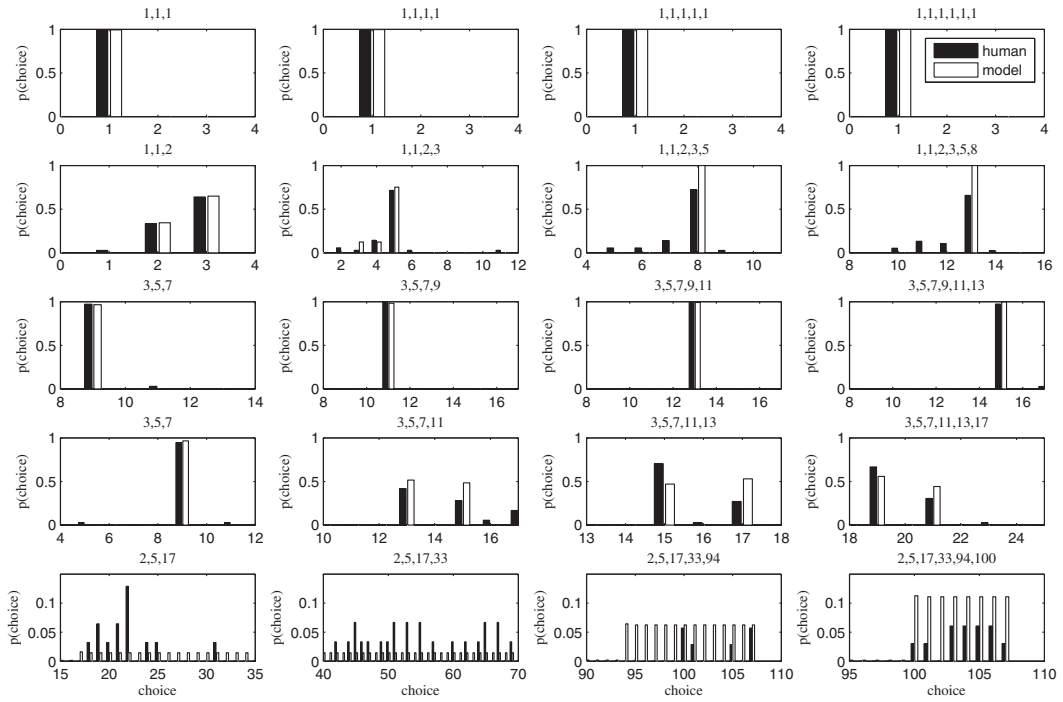

Fig. 1. Results of Experiment 1. Each row of plots shows the predictions for one sequence as the number of elements increases from 3 to 7 across the columns. The five rules used to generate the sequences are (from top to bottom) repeating ones, sum of the last two numbers, increasing odd numbers, increasing odd prime numbers, and increasing numbers. Note that the scale of the increasing numbers is different from the other plots, and it may omit some values of both distributions for visual clarity.

.004 for ''repeat the last k_h numbers,'' .0249 for ''kth power,'' .0001 for ''prime numbers,'' and .2499 for the set of stochastic rules.[4]

Although not essential for applying our theoretical analysis, the quantitative fit of the Bayesian model to human predictions is good. The correlation between the sets of predictions is $r = .88$. As the ''increasing numbers'' pattern is unpredictably random, both the participant and model predictive distributions are diffuse, lowering this correlation. The predictive distributions are nearly identical for the four deterministic sequences, with $r = .98$. If we remove the sequences for which people and the model only produce a single prediction (the first and third deterministic sequences), the correlation between responses is $r = .80$ (this includes the ''increasing numbers'' pattern). As the correlation between the participant and model predictions is increased every time both produce near-zero probability for a number, we also calculated a more stringent G^2 test statistic between the responses. The result of the G^2 test is that the model and human responses are not significantly different ($\chi^2(2194) = 838.87, p = 1.0$).

To demonstrate that the correlations are robust to variations in the parameter values, we looked at the model's sensitivity to random variations in the parameter values. To do this,

we looked at the quantitative fit of the Bayesian model with randomly drawn parameters $\theta \sim \text{Dirichlet}(\alpha\theta^*)$, where θ^* are the parameters fit to the data and α controls how much we vary from the fitted parameters. We chose $\alpha =$ 50 and 1,000 (which yield standard deviations in the parameter for the most probable rule of seven and two percent, respectively) and calculated the correlation between the predictive distributions for 20 randomly drawn parameters. Including all sequences, $r = .87$ with a standard deviation of 0.004 for $\alpha = 50$ and $r = .88$ with a standard deviation of 0.0004 for $\alpha = 1{,}000$. Considering only the deterministic sequences, $r = 0.98$ with a standard deviation of 0.005 for $\alpha = 50$ and $r = .98$ with a standard deviation of 0.0005 for $\alpha = 1{,}000$. Finally the correlations removing the sequences with a single prediction are $r = .79$ with a standard deviation of 0.007 for $\alpha = 50$ and $r = .80$ with a standard deviation of 0.0007 for $\alpha = 1{,}000$.

Even though the quantitative fit of the model is good, there is still some discrepancy between human and model predictions. For example, there are some clear differences between the two distributions for the ''prime numbers'' and Fibonacci rules. This could be due to the fact that many participants who did not know the prime or Fibonacci numbers would invent a complex rule composed of multiple ''$\times C + K$'' rules, which was not included in the model (e.g., +2,+2,+4,+2,+2,… for the sequence [3,5,7,11,13]). Thus, the current model cannot capture this aspect of human behavior.

As future work, it could be interesting to extending the model using a compositional rule grammar to generate hypotheses could explain the combinations of $\times C + K$ that many participants invent, as was done by Coen and Gao (2009). One example of how this extension to the simple Bayesian model could explain cases where human responses deviate from the model's predictions is the Fibonacci sequence (sequence 2 of Experiment 1). For example, the second most predicted number for the sequence (1,1,2,3,5) is seven, which would be predicted by the following simple composition of $\times C + K$ rules: +0,+1,+2,+1,+0…. We consider how other discrepancies between the Bayesian model and human responses could be explained at the process level in Section 8.

5. Experiment 2: Predicting probability ratings

The results of Experiment 1 demonstrate that the model can describe human predictions on simple number sequences. In addition to capturing these predictions, we would like our model to have similar uncertainty to people in predicting which number will be next. This way, the model would also capture how much information is gained by learning the next number of the sequence, giving us an indication of how similar its assessment of expected information gain might be to those of our participants. In Experiment 1, we asked each participant to provide a single estimate of which number they believed will be next for a given sequence. A single estimate does not capture how certain participants are of which number is next. Thus, in Experiment 2 we asked participants to rate the probability that each of four different numbers would be the next number in the sequence. We then used these ratings to compare levels of uncertainty between participants and our model.

5.1. Method

5.1.1. Participants

Participants were 20 volunteers from the undergraduate population at Brown University.

5.1.2. Stimuli

The four sequences tested were (1,1,1,1,1), (2,4,6,7,8,10), (1,1,2,3,5,8), and (3,5,7,9,11,13). Four possibilities for the next number were chosen for each subsequence of these sequences. Possible next numbers were selected to be consistent with multiple rules matching each subsequence.

5.1.3. Procedure

The experiment was presented as a written survey. The participants received the following instructions: ''I have created four sequences where the numbers are related by a simple relational rule, such as increasing numbers, decreasing numbers, increasing odd numbers, decreasing even numbers, and the last number plus one.'' The survey began with the first number of a sequence and asked the participants to rate how likely (from 1 to 7) they feel four different numbers are to be the next number in the sequence. After they finished rating the four potential next numbers, the next number in the sequence was given and four new possible next numbers to rate. This repeated until the participants had received the full sequence, at which point the next sequence began.

5.2. Results and discussion

To get the probability of each next number from the judgments produced by the participants, we subtracted one from each rating, averaged over participants, and normalized over the four judgments for the sequences at each length. To compare our model to human judgments, we computed the probability of each next number according to our model using Eqs. 1 and 2 and the hypothesis space defined above. We used the parameters established in Experiment 1, allowing us to test whether these parameters overfit the data from that experiment. Fig. 2 shows the aggregate human and model predictive distributions for the sequences at all lengths. The correlation between our model using these parameters and participant judgments was $r = .68$, providing support for the claim that our model is not overfitting the data and is robust to technique as in the previous experiment, the correlation between the predictive distributions for 20 randomly drawn parameters was $r = .67$ with a standard deviation of 0.0093 for $\alpha = 50$ and was $r = .68$ with a standard deviation of 0.0008 for $\alpha = 1{,}000$.

Analysis of the entropy of the distributions produced by the model and people suggested that both are similarly surprised by the same next numbers. In other words, the change in the entropy in the model and human predictive distributions is qualitively similar. Aside from (1,1,2,3), both the model and human show the same qualitative changes in entropy. For example, as the sequence (2,4,6) grows from (2) to (2,4,6), both participants and the model become more certain that the rule is ''+2.'' As 7 is the next

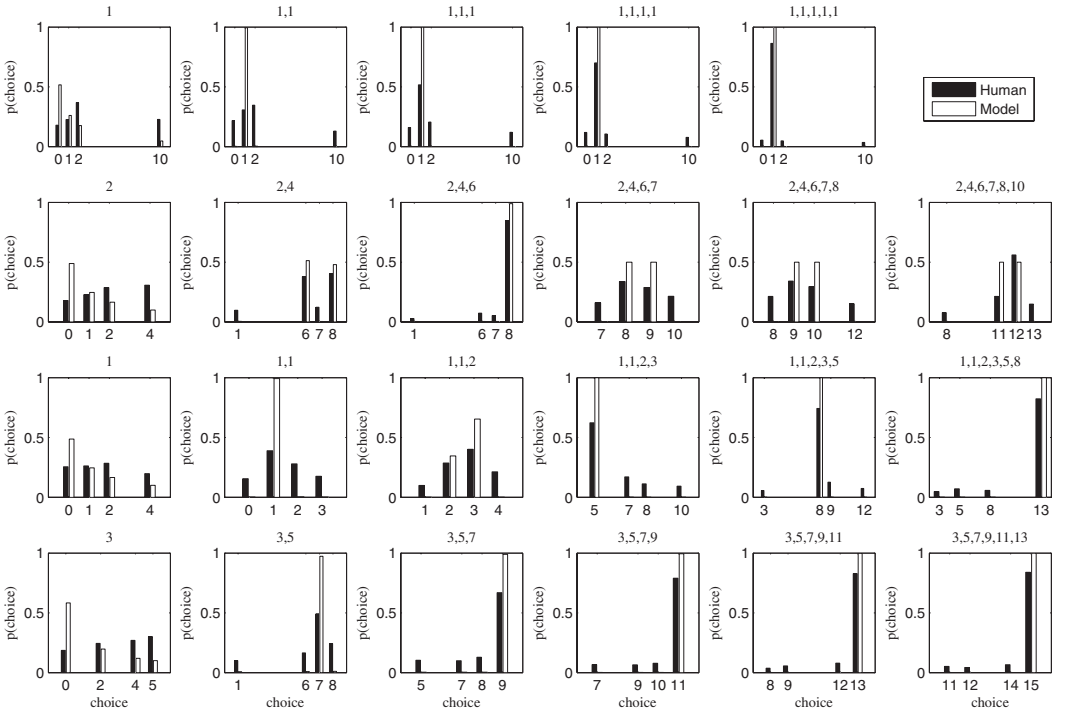

Fig. 2. Results of Experiment 2. Each column of plots shows the predictions for one sequence as the number of elements increases from 1 to 6 (5 for the repeating ones sequence).

number, the previously most likely rule is ''+2'' is falsified and there is no clear deterministic alternate. Thus, the entropy of both distributions appropriately increases. The expected information gain thus seems to be roughly in accordance between people and the model.

Although the judgments are qualitatively similar, the model clearly does not perfectly capture human predictive judgments. Using the more stringent G^2 test, the two distributions are statistically significantly different ($\chi^2(69) = 1470, p < .001$). One clear issue is that people are much more conservative in their judgments on this task. This is echoed in the fact that the human entropy is always larger than the model entropy (see Fig. 3). This is not a major concern for our analysis for two main reasons. First, we are not proposing this as a model of how people predict the next number in a sequence, but rather as a tool for testing the predictions of our theoretical analysis. It is clear from the first two experiments that the Bayesian model's predictions are similar to human predictions. Second, it is possible that the difference between people and the model's predictions is due to a process-level factor. For example, people may be acting more conservatively because they think there may be an aspect of deception in the experiment or because rating how likely different numbers will be the next number is an unnatural task. These factors are not taken into account in our more abstract, computational-level model.

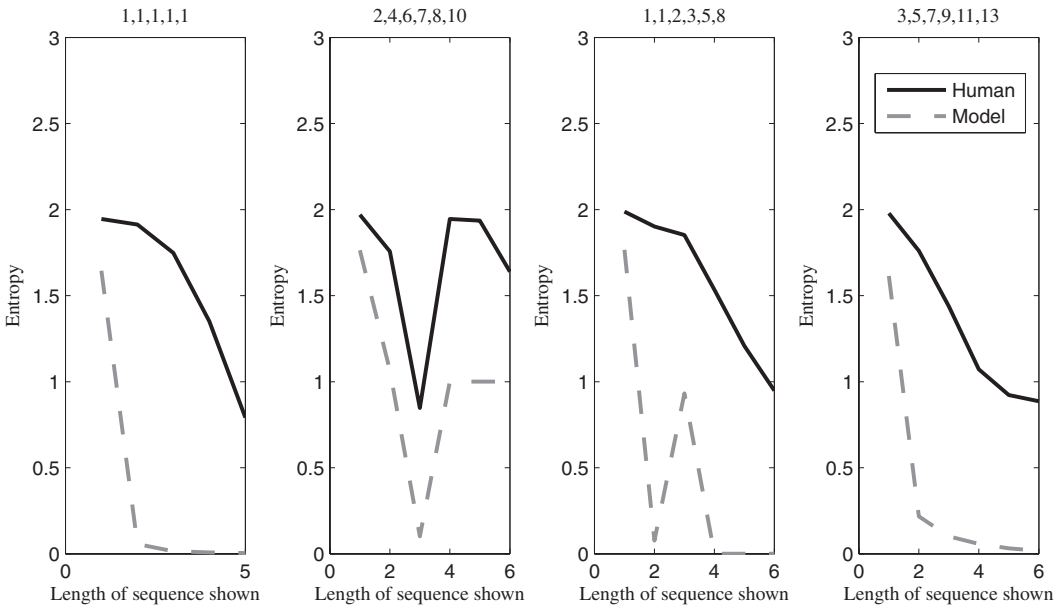

Fig. 3. The change in entropy as each sequence increases in length for both the human aggregate and model predictive distributions for Experiment 2.

6. Experiment 3: Testing confirmation of the most probable hypothesis

Having verified that a Bayesian model can capture human sequence predictions, we can use it to test how human hypothesis testing is affected by prior knowledge. Our analysis of optimal hypothesis testing predicts that people should seek to confirm the hypothesis that they assign highest posterior probability. To test this prediction, Experiment 3 manipulated the prior probability of different types of hypotheses to see whether we could induce people to change which hypotheses they sought to confirm.

6.1. Methods

6.1.1. Participants

A total of 67 undergraduates from the University of California, Berkeley participated in exchange for course credit. Participants were split into three conditions, with 22 participants in the $\times C + K$ condition, 22 participants in the *sum last two* condition, and 23 participants in the *control* condition.

6.1.2. Stimuli

In order to establish the priors in different sequence prediction environments, participants in the $\times C + K$ and *sum last two* conditions were trained on 100 sequences of numbers. The training sequences in the $\times C + K$ condition had a high prevalence (87%) of sequences

generated by rules of the form ''$\times C + K$'' and no sequences generated by summing the last two numbers, and vice versa in the *sum last two* condition (with 89% of sequences conforming to the ''sum of the last two numbers'' rule). No training was provided in the *control* condition. Test selection was probed with 21 sequences consistent with both the ''sum of the last two numbers'' and the ''$\times C + K$'' rules, shown to participants in all three conditions. For example, one of the 21 test sequences, (3,6,9), can be interpreted as $\times 1 + 3$ or the sum of the last two numbers ($3 + 6 = 9$).

6.1.3. Procedure

The experiment had two phases: a training phase and a test phase. In the training phase, participants were asked to predict the next number in the sequence and the underlying rule, and then told whether their responses were correct. The group of participants in the *control* condition were not trained on any sequences and only were given the test portion of the experiment. In the test phase, participants were told that they could pick one number and find out whether that number was the next in the sequence, being told to select the number that would help them figure out the underlying rule the best. They were asked to write down both what they thought the rule was and their number choice. The experiment was administered on a computer with instructions given by the experimenter. The participants were also provided a calculator to ensure that arithmetic ability was not a factor in people's responses.

6.2. Results and discussion

If participants are sensitive to the prior probabilities of different environments, then they should choose to confirm the same rule as their training condition. As the priors in both the *control* (established by the priors learned from Experiment 1) and $\times C + K$ conditions are similar, our main concern is whether participants are more likely to confirm the ''sum of the last two numbers'' rule when trained in the *sum last two* condition. For all of the test sequences, the model predicts confirmation of the current hypothesis, which in turn is determined by the prior probabilities established by the training condition.

The responses produced by the participants for all sequences were grouped into three categories: $\times C + K$, sum of the last two numbers, or other. Two coders who were blind to condition assigned the rules people selected as belonging to these three groups, with high inter-rater reliability ($\kappa = 0.90$). Disagreements were resolved through discussion. As the model predicts, participants were sensitive to the environment given in their training condition and changed their responses appropriately (see Fig. 4). Although participants did not confirm the appropriate hypothesis for every sequence as the model predicts, the variation was statistically significant. Participants in the *sum last two* condition tested the ''sum of the last two numbers'' rule significantly more often than participants in either the $\times C + K$ condition ($\chi^2(2) = 9.71, p < .01$) or the *control* condition ($\chi^2(2) = 9.35, p < .01$). Additionally, the responses for the $\times C + K$ and *control* conditions were not significantly different ($\chi^2(2) = 1.11,\ p > .55$). Thus, when testing their theories and hypotheses, people are sensitive to the prior probabilities in the environment, choosing to confirm the hypothesis rendered most probable by that environment.

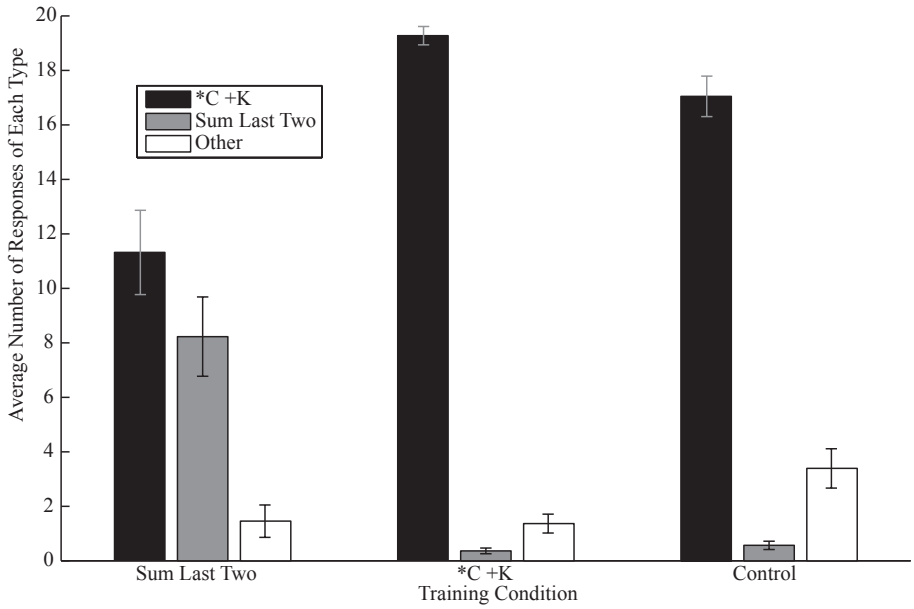

Fig. 4. Results of Experiment 3, averaged over participants in each group. Error bars show one standard error.

7. Experiment 4: Predicting or testing?

The results of Experiment 3 suggest that participants are sensitive to the base rates of rules in their environment when predicting the next event in a sequence and performing tests to discover the rule producing a sequence. However, according to our theoretical analysis, both the most likely next number in the sequence and the optimal test are the same number. Thus, it is not clear whether participants in the second part of Experiment 3 actually processed their instructions to perform tests to discover the rule, or simply guessed the next number in the sequence (like they did in the first part of the experiments). To rule out this alternative explanation, Experiment 4 replicated Experiment 3 except the training phase was replaced with a feedback phase, where participants performed tests on the training sequences from Experiment 3 and were given feedback as to whether or not their tests were successful. As participants are actively testing throughout the entire experiment and are never asked to predict the next number of a sequence, the alternative explanation that they are making predictions during the test phase does not apply.

7.1. Methods

7.1.1. Participants

A total of 29 undergraduates from the University of California, Berkeley participated in exchange for course credit. Participants were split into two conditions, with 12 participants in the $\times C + K$ condition and 13 participants in the *sum last two* condition. Four participants

wrote the number to test twice instead of the most likely hypothesis to have generated the data and were removed from further analysis.

7.1.2. Stimuli

The training sequences for the $\times C + K$ and *sum last two* conditions from Experiment 3 were used in Experiment 4. The same set of test sequences from Experiment 3 was used in both conditions of Experiment 4.

7.1.3. Procedure

The experiment had two phases: a feedback phase and a test phase. In the feedback phase, participants were given the task of discovering the underlying rule of a sequence for 100 sequences. They were told that to achieve the goal, they got to choose a number and they would receive feedback as to whether or not it is the next number in the sequence. They were reminded that more than one rule could fit the sequence they saw, so that they should pick the number that helps them figure out the underlying rule as best as possible. For the *test* phase, participants were told that it was identical to the first part, except that they would not be given feedback, but that they should still make their choices as if it were given. They were asked to write down both what they thought the rule was and their number choice. The experiment was administered on a computer with instructions read by an experimenter. The participants were also provided a calculator to ensure that arithmetic ability was not a factor in people's responses.

7.2. Results and discussion

Fig. 5 shows the average numbers of tests of each type made by participants in Experiment 4. Replicating Experiment 3, participants adapt their tests appropriately to their environment. Participants in the *sum last two* condition test the *sum last two* rule more often than those in the $\times C + K$ condition ($\chi^2(2) = 8.14$, $p < .05$). As participants only performed tests and never predicted the next number in the experiment, the alternative explanation that participants in Experiment 3 continued predicting in the test phase has been ruled out. The combination of Experiments 3 and 4 provide strong evidence that participants adapt their testing strategy in accordance with the prior probability of rules in their environment in the direction predicted by our analysis.

8. General discussion

We have shown that the PTS is optimal under the assumption that the hypotheses under consideration are deterministic, using both maximizing the probability of falsification and reduction of uncertainty as measures of test utility. Our experiments provide the pieces of evidence needed to connect this result to human behavior. Experiment 1 showed that a Bayesian model of sequence prediction accurately characterizes human expectations. Experiment 2 showed that this model could also capture levels of

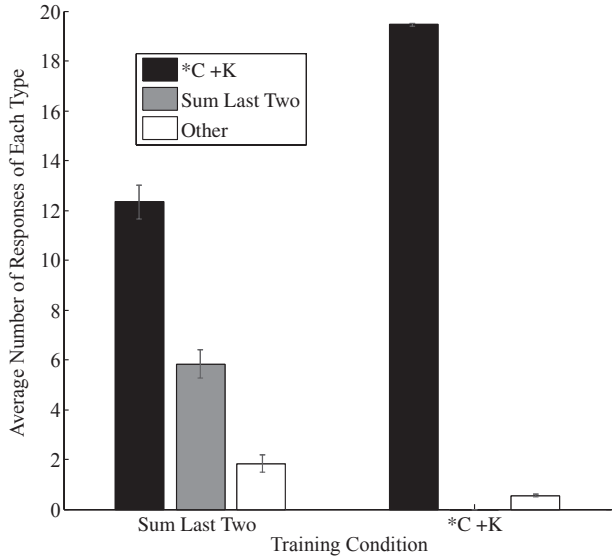

Fig. 5. Results of Experiment 4, averaged over participants in each group. Error bars show one standard error. These results replicate Experiment 3 and demonstrate participants are actively testing, and not just predicting the next number.

uncertainty, and how this uncertainty changed with new observations. These experiments thus supported our use of a Bayesian framework in analyzing how people test their hypotheses. Our formal analysis predicts that changing the relative prior probabilities of two hypotheses that could both have generated ambiguous sequences should change the test that people choose. Experiments 3 and 4 showed that people behave in a way that matches this prediction, selecting tests that confirmed the hypothesis most probable in each environment. Thus, participants are not blindly testing the same choice regardless of the environment, but are identifying the most probable hypothesis and then systematically seeking to confirm that hypothesis. In this section, we consider the plausibility of the assumption of determinism, how these results relate to previous work on the confirmation bias, and close by briefly summarizing their implications for understanding why people might exhibit such a bias.

8.1. How realistic is the assumption of determinism?

Our analysis shows that the positive test strategy can be optimal when the hypotheses being evaluated are deterministic. This suggests a natural explanation for why people might pursue this strategy: They assume that the kinds of causal systems they encounter operate deterministically. However, the plausibility of this explanation depends on the existence of independent evidence that people make this assumption. Fortunately, a number of recent studies suggest that both adults and children seem to operate under the assumption that causal relationships are deterministic.

In discussing determinism, it is useful to distinguish between ''weak'' determinism—the idea that all events have causes—and ''strong'' determinism—the idea that causes produce their effects every time (e.g., Schulz & Sommerville, 2006). Evidence that people minimally assume weak determinism comes from a variety of studies in causal learning and reasoning, but perhaps the most systematic characterization of this assumption appears in the literature on psychological essentialism (Gelman, 2003; Gelman et al., 1994). Psychological essentialism asserts that hidden behind every object or event we observe is something underlying it (its ''essence'') that explains its existence. The essence usually is causally responsible for the observable properties of the object or event. Evidence that people hold such a belief comes from studies showing that children search for hidden causal mechanisms when an event violates known causal laws (Chandler & Lalonde, 1994) and adults (Luhmann & Ahn, 2003).

Our analysis assumes the stronger form of determinism, with the probability of the next event in a sequence being either zero or one. Evidence that children make this kind of assumption comes from recent work by Schulz and Sommerville (2006). A series of experiments demonstrated that when children observe a stochastic cause, they resist accepting that it is inherently stochastic. Instead, they infer either unobserved inhibitory or generative causes appropriate to the probabilistic information given (Schulz & Sommerville, 2006). As this is the determinism assumption used in our analysis, this provides empirical justification that children do not believe rules are inherently probabilistic—even when provided probabilistic information. Further experiments with adults suggest that the assumption of determinism is necessary in order to explain how people learn certain kinds of causal relationships so quickly, and that this assumption operates as a default in certain domains (Griffiths, 2005; T. L. Griffiths, D. M. Sobel, J. B. Tenenbaum, & A. Gopnik, unpublished data).

Intuitively, the strong form of determinism is appealing. Most of the laws in natural domains follow strong determinism (e.g., gravity in intuitive physics). Although intuitively appealing, clearly further research is needed to determine how domain specific the assumption of the strong form of determinism is, as previous experiments focused on mechanical relations. In particular, establishing whether people assume that mathematical rules are deterministic seems necessary in order to understand the difficulty that people have with falsification in contexts like that of Wason's (1960) 2-4-6 task (the low prior probability assigned to mixtures of ''$\times C + K$'' rules in our experiments provides preliminary support for this idea). Our results demonstrate the importance of exploring the psychological validity and the problems with this assumption; if we do actually assume strong determinism, our tendency to use the PTS is rational, at least relative to our assumptions about the world.

8.2. Relationship to previous work on the PTS

Though our analytic results are a novel justification of our tendency to use the PTS, other researchers have also provided mathematical arguments that the PTS can be a sensible strategy under certain circumstances. Klayman and Ha (1987) derived general conditions for when the probability of falsifying the current hypothesis with the PTS is greater than with

the NTS. They demonstrated that the PTS is often optimal when the number of events that the true rule applies to is small (the minority phenomenon or rarity assumption), a result that is in concordance with Oaksford and Chater's (1994) analysis of confirmation in the card selection task. A similar result was recently presented by Perfors and Navarro (2009). Additionally, Klayman (1995) argued that the PTS is justified if one is only interested in sufficiency and not necessity. In other words, if one is only interested in a hypothesis that can explain all of the events one has observed but perhaps not the only hypothesis that can do so, then testing events true under the candidate hypothesis is sensible. Our work complements these analyses by providing another justification for the human bias toward the PTS, founded in an assumption about the structure of the environment. Interesting directions for further research include exploring how the optimality of the PTS is affected by incorporating all of these assumptions together, and how robust it is to mild violations of these assumptions (such as including a few low-probability nondeterministic hypotheses, as we did by allowing mixtures of deterministic rules in our experiments).

Klayman and Ha (1987) proposed exploring the role of the set of possible hypotheses on testing; however, few papers have used constrained hypothesis spaces to analyze the optimality of different strategies. One exception is Nelson and Movellan (2001) who explored directly applying the Bayesian generalization model of Tenenbaum (1999) and Tenenbaum and Griffiths (2001) to a task similar to Wason's (1960) 2-4-6 task, using EIG to determine optimal tests. In their task, hypotheses were sets of numbers and the goal was to find the hypothesis most likely to have generated a given set of numbers. The participants were allowed to ask whether one other number followed the rule. Nelson and Movellan (2001) found that in cases of high posterior uncertainty, the tests predicted by EIG matched the tests given by simply seeking confirmation; however, in cases of low uncertainty, the tests predicted by EIG conflicted with the tests given by seeking confirmation (and with those selected by human participants). One representive example where human responses deviate from EIG is for the given set {60,80,10,30}. Here, the working hypothesis is multiples of 10, but multiples of five is also possible. In this case, analogous to the original 2-4-6 task, the alternative hypothesis (increasing numbers for Wason (1960), multiples of five for Nelson and Movellan (2001) picks a superset of the outcomes consistent with the most probable hypothesis. This is where our analysis differs from previous work: By assuming that hypotheses are deterministic, we require them to pick only a single prediction and thus no hypothesis strictly subsumes another.

8.3. Connecting to process models of sequence prediction and testing

The Bayesian model of sequence prediction that we developed in this paper was intended primarily to allow us to explore whether the assumptions behind our analysis of hypothesis-testing strategies held with human learners. This model was specified at Marr's (1982) computational level, focusing on the abstract problem of sequence prediction and its solution in terms of Bayesian inference. Using this model, we could examine whether human behavior approximated Bayesian inference and estimate prior probabilities for different hypotheses that then allowed us to manipulate these probabilities and examine the effects on hypothesis

testing. However, it is still an open question how people are solving this problem at the algorithmic level, with the development of appropriate psychological process models being an interesting direction for future work.

As with many Bayesian models, it is clear people are not explicitly performing exact Bayesian inference when they are solving problems of sequence prediction and hypothesis selection. In other words, people are not creating a giant list of all the possible hypotheses in their minds and summing the result of multiplying the prior and likelihoods for each hypothesis. Additionally, people are not explicitly using a giant list of posterior probabilities to calculate the EIG or probability of falsifying using PTS or NTS. However, our experiments do suggest that people are changing their beliefs in a way that is approximately consistent with Bayes' rule, and changing their testing behavior in accordance with the prior probabilities of the rules they observe in the environment. These results provide a constraint on process models, indicating that whatever the underlying psychological mechanisms are they need to yield behavior that is approximately Bayesian.

One way to define process models that satisfy this constraint is to start with methods of approximating Bayesian inference that have been developed in computer science and statistics, and then consider how those methods could be implemented in psychologically plausible ways. This is the strategy that is taken in deriving ''rational'' process models (Sanborn et al., 2006; Shi, Feldman, & Griffiths, 2008), which approximate Bayesian inference under certain processing constraints (such as limits on the amount of memory or computation that can be used). Many of these models are based on the Monte Carlo principle, approximating a probability distribution with a set of samples from that distribution. In particular, the problem of updating a probability distribution over hypotheses as more observations are acquired (which arises in sequence prediction) can be approximated using a particle filter, in which the distribution at each stage is approximated by a set of samples (known as ''particles'') (Doucet, Freitas, & Gordon, 2001). Particle filters have become a popular method for exploring the effects of processing constraints in probabilistic models (Brown & Steyvers, 2009; Levy, Reali, & Griffiths, 2009; Sanborn et al., 2006), and they may provide a way of constructing more psychologically plausible models of sequence prediction.

9. Conclusion

Since Wason's (1960) introduction of the 2-4-6 task, philosophers and psychologists have been fascinated by explaining why people tend to test their hypotheses about the world with examples that are true under their hypothesis. We have provided a novel explanation for the effect: that this strategy can be optimal, if people assume the rules underlying the world are deterministic. Once this assumption is made, choosing examples that are true under their current hypothesis provides a rational way to test that hypothesis, whether people's normative standard is to maximize their probability of falsifying their current hypothesis or the expected information they will gain through their test. Our experiments support the basic assumptions behind this analysis, showing that people make predictions in a way that can be

characterized by Bayesian inference and that their selection of which tests to perform is sensitive to the posterior probability of different hypotheses. Furthermore, recent studies of human causal learning suggest that people often assume the kind of determinism required by our analysis. Taken together, these results suggest that the assumption of determinism can play a similar role in explaining why people might pursue a strategy of confirmation when predicting sequences to the role that rarity plays in explaining how people choose to test logical rules (Oaksford & Chater, 1994): While people might not always pursue the best strategy for the problem posed by the experimenter, they act in a way that is rational with respect to their assumptions about their environment.

Notes

1. More precisely, choosing the hypothesis with highest posterior probability is always at least as good as choosing any other hypothesis, with equality holding in the case where just two hypotheses have nonzero posterior probability.
2. In presenting the model, we sometimes use process-level language in order to make the mathematics more intuitive, but this should not be interpreted as a claim that the mathematical operations we describe are intended to be interpreted as psychological processes.
3. The different penalizations from the initial number distribution based on k_h does not have a unique effect on the model predictions because its effect is confounded with the parameters from the prior probability.
4. The estimates of the prior probabilities were based on a relatively large amount of data—by allowing freeform responses we gained a lot of information about the values of the parameters from each response—and we were thus not concerned about overfitting the data. As our goal was primarily to determine whether our model incorporated an appropriate set of hypotheses—something that could be seen even allowing a free parameter for each rule type—the fact that the model has six free parameters did not interfere with evaluating its performance. As a further check against overfitting, we test the generalization performance of the model in Experiment 2.

Acknowledgments

This project was initiated while the authors were in the Department of Cognitive and Linguistic Sciences at Brown University. A preliminary version of Experiments 1 and 3 was presented at the 30th Annual Meeting of the Cognitive Science Society. We thank David McNamee, Kevin Canini, and the UC Berkeley Computational Cognitive Science Lab for thoughtful discussions, our research assistants for help running experiments, and Mike Byrne, Mike Oaksford, and three anonymous reviewers for valuable comments on previous drafts of this manuscript. This work was supported by grant FA9550-07-1-0351 from the Air Force Office of Scientific Research.

References

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*, 49–67.

Chandler, M. J., & Lalonde, C. E. (1994). Surprising, miraculous, and magical turns of events. *British Journal of Developmental Psychology*, *12*, 83–95.

Coen, M. H., & Gao, Y. (2009). Learning from games: Inductive bias and Bayesian inference. In N. Taatgen & H. van Rijk (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 2729–2734). Austin, TX: Cognitive Science Society.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Doucet, A., de Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. New York: Springer.

Gelman, S. (2003). *The essential child*. Oxford, England: Oxford University Press.

Gelman, S., Coley, J. D., & Gottfried, G. M. (1994). Essentialist beliefs in children: The acquisition of concepts and theories. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 341–365). Cambridge, England: Cambridge University Press.

Griffiths, T. L. (2005). *Causes, coincidences, and theories*. Unpublished doctoral dissertation, Stanford University.

Klayman, J. (1987). *An information theory analysis of the value of information in hypothesis testing* (Tech. Rep. No. 119a). Chicago: University of Chicago.

Klayman, J. (1995). Varieties of confirmation bias. In J. Busemeyer, R. Hastie & D. Medin (Eds.), *Psychology of learning and motivation: Vol. 32. Decision making from a cognitive perspective* (pp. 365–418). New York: Academic Press.

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

Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21* (pp. 937–944). Cambridge, MA: MIT Press.

Luhmann, C. C., & Ahn, W. (2003). Evaluating the causal role of unobserved variables. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 24th annual conference of the Cognitive Science Society* (pp. 734–739). Boston, MA: Cognitive Science Society.

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

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*, 979–999.

Nelson, J. D., & Movellan, J. R. (2001). Active inference in concept learning. *Advances in Neural Information Processing Systems*, *13*, 45–51.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.

Oaksford, M., & Chater, N., Grainger, B., & Larkin, J. (1997). Optimal data selection in the reduced array selection task (rast). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(2), 441–458.

Perfors, A. F., & Navarro, D. J. (2009). Confirmation bias is rational when hypotheses are sparse. In N. Taatgen & H. van Rijk (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 2741–2746). Austin, TX: Cognitive Science Society.

Popper, K. R. (1935/1990). *The logic of scientific discovery*. Boston: Unwin Hyman.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the Cognitive Science Society* (pp. 726–731). Mahwah, NJ: Erlbaum.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschool causal inferences. *Child Development*, *77*(2), 427–442.

Shannon, C. E. (1948). The mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Shi, L., Feldman, N. H., & Griffiths, T. L. (2008). Performing Bayesian inference with exemplar models. In B. Love, K. McRae, & V. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 745–750). Austin, TX: Cognitive Science Society.

Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–641.

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

Wason, P. C. (1968a). On the failure to eliminate hypotheses – a second look. In P. C. Wason & P. N. Johnson-Laird (Eds.), *Thinking and reasoning* (pp. 165–174). Harmondsworth, Middlesex, England: Penguin.

Wason, P. C. (1968b). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.