A Rational Analysis of Confirmation with Deterministic Hypotheses

Joseph L. Austerweil (Joseph.Austerweil@gmail.com)
Thomas L. Griffiths (Tom_Griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, Berkeley, CA 94720-1650 USA

Abstract

Whether scientists test their hypotheses as they ought to has interested both cognitive psychologists and philosophers of science. Classic analyses of hypothesis testing assume that people should pick the test with the largest probability of falsifying their current hypothesis, while experiments have shown that people tend to select tests consistent with that hypothesis. Using two different normative standards, we prove that seeking evidence predicted by your current hypothesis is optimal when the hypotheses in question are deterministic and other reasonable assumptions hold. We test this account with two experiments using a sequential prediction task, in which people guess the next number in a sequence. Experiment 1 shows that people's predictions can be captured by a simple Bayesian model. Experiment 2 manipulates people's beliefs about the probabilities of different hypotheses, and shows that they confirm whichever hypothesis they are led to believe is most likely.

Keywords: confirmation bias; rational analysis; hypothesis testing; Bayesian inference.

How *should* a scientist seek evidence to help her find the hypothesis that explains a phenomenon? Does this differ from how people *do* seek evidence? Popper (1935/1990) argued that scientists ought to follow the strategy of *falsification*, seeking evidence most likely to falsify their current theory. Interested in whether people adhere to this strategy, Wason (1960) investigated how people intuitively test their theories. In the classic 2-4-6 task, participants were asked to uncover a relational rule 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 a potential rule also being valid under the increasing numbers rule. Thus, the true rule can only be discovered by testing numbers that are not predicted by your current best guess at the rule (negative test strategy or NTS). Rather than follow the NTS, participants choose to test triplets predicted by their current hypothesis (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* or 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. Previous work has identified settings in which 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 might explain 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., Schulz & Sommerville, 2006; Gelman, Coley, & Gottfried, 1994). This emphasis on the structure of the environment parallels similar strategies pursued in other rational analyses (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 a behavioral experiment 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. In a second experiment, we demonstrate that changing a person's beliefs about the probability of hypotheses affects their evidence-seeking strategy. We conclude by discussing how our results relate to previous work.

Sequence prediction and hypothesis testing

Given a sequence of events, how do we predict what will occur next? For example, suppose you see a woman outside of an airport and then at the security checkpoint. How likely is it that she stays at the security checkpoint (she is a security guard) or walks to a gate (she is a passenger or crewmember)? 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. Since there is no means of predicting the next event with complete certainty, this is an inductive task.

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

$$P(x_i|\vec{x}) = \sum_h P(x_i|h,\vec{x})P(h|\vec{x}) \quad (1)$$

where $P(x_i|h,\vec{x})$ is the probability of the next event under hypothesis h, and $P(h|\vec{x})$ is the posterior probability of that hypothesis given the sequence $\vec{x}$. This posterior probability can

be obtained from Bayes' rule, with

$$P(h|\vec{x}) = \frac{P(\vec{x}|h)P(h)}{\sum_{h'} P(\vec{x}|h')P(h')} \quad (2)$$

being the normalized product of the likelihood, $P(\vec{x}|h)$, 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 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 determing the best question to test her hypothesis. In the remainder of the section, we show that there is a simple answer to this question provided our hypotheses are *deterministic*, allowing only one value for x given $\vec{x}$ (ie., that there is only one place the woman will go for each hypothesis about her identity). In this case, the positive test strategy (asking about the event that corresponds to the most probable hypothesis) is optimal. Thus, the best question is to ask her husband is whether she will be in the location that our best guess about her identity predicts.

We will use two methods to identify what question we should ask. The first is the probability of falsification - asking the qeustion that gives us the highest probability of falsifying our current hypothesis (Popper, 1935/1990; Klayman & Ha, 1987). The second is a measure based on information theory (Klayman, 1987; Oaksford & Chater, 1994). According to information theory, 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 $P(h|\vec{x})$, given a sequence of objects from performing a test c and learning its outcome r, is

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

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

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

with

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

being the probability of the outcome r from the test c given our previous observations $\vec{x}$. In sequential 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 $\vec{x}$.

Since 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* (EIG). The optimal choice $\hat{c}$ is

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

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

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

being that which minimizes uncertainty after the response.

The optimality of positive test strategies

Instead of directly deriving general results on the usefulness of positive test strategies, we first consider the problem with simplifying assumptions. We narrow our hypothesis space to deterministic hypotheses which all make different predictions for the next event in the observed sequence. 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 the probability of falsifying that hypothesis and the EIG.

Using maximizing the probability of falsifying the current working hypothesis as our normative standard (Klayman & Ha, 1987), the analysis is simple. The probability that testing the choice c, consistent with hypothesis h^c, falsifies that hypothesis is $1 - P(h^c|\vec{x})$. If you want to falsify a particular hypothesis, then it is best to test the choice it predicts since $1 - P(h^c|\vec{x})$ is the sum of the probabilities of all other hypotheses. Consequently, to falsify the current hypothesis, you should test the choice it predicts and thus the PTS is optimal.

The same result holds when we take maximizing the EIG 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 events next, 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

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

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

$$-(1-P(h^c|\vec{x}))\sum_{h\neq h^c}\frac{P(h|\vec{x})}{1-P(h^c|\vec{x})}\log_2\frac{P(h|\vec{x})}{1-P(h^c|\vec{x})}$$
$$=-\sum_{h\neq h^c}P(h|\vec{x})\log_2 P(h|\vec{x})+\sum_{h\neq h^c}P(h|\vec{x})\log_2(1-P(h^c|\vec{x}))$$

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

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

Since 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

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

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 0.5. It is easy to show that this is the hypothesis with highest posterior probability.[1] There are two cases. If all probabilities $P(h|\vec{x})$ are less than 0.5, then the hypothesis for which $P(h|\vec{x})$ is greatest is clearly the closest to 0.5. If the probability of some hypothesis is greater than 0.5 there is only one such hypothesis, and the distance of the probability of all other hypotheses from 0.5 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.

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 $\mathcal{H}^c$ be the set of hypotheses that predict c as the next event and $\mathcal{H}^{\bar{c}}$ 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 $\mathcal{H}^c$, with $P(\mathcal{H}^c|\vec{x})=\sum_{h\in\mathcal{H}^c}P(h|\vec{x})$, although there is an extra wrinkle produced by the fact that confirmation does not guarantee an entropy of 0. This analysis shows that the optimal test is that which produces $P(\mathcal{H}^c|\vec{x})$ closest to 0.5. 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 0.5. 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 positive test strategy is optimal.

[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 non-zero posterior probability.

A Bayesian model for 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 domain of our model of sequence prediction is numbers. We assume that the sequence of observed numbers, $\vec{x}=(x_1,\ldots,x_{i-1})$, is generated from some relational rule $h\rightarrow\vec{x}$, and that people try to identify this rule in order to make accurate predictions. Our 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. 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 be a reasonable approximation that we can use to understand human hypothesis testing.

Instead of defining the hypotheses as sets of numbers, each hypothesis is a rule from k_h previous numbers to the possible next numbers of the sequence. The likelihood assigns a probability distribution over next numbers given the previous k_h observed numbers. We divide the types of hypotheses into two separate categories: deterministic and non-deterministic. A *deterministic* hypothesis, such as increasing odd numbers, has only one correct next number and conforms to the following form: $h(x_{i-1},\ldots,x_{i-k+1})\colon X^k\rightarrow X$. 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, more than one number may conform to a *non-deterministic* hypothesis. For example, the following likeli-

hood function models the increasing numbers ($k_h = 1$):

$$P(x_i|h, x_{i-1}; \nu) = \begin{cases} \frac{1}{\nu+1} & x_i \geq x_{i-1} \wedge x_i - x_{i-1} \leq \nu \\ 0 & x_i < x_{i-1} \end{cases}$$

where ν is the largest increase possible from the last number.

These hypotheses 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, repeat the last k_h numbers, the i-th power (for $i = 2$ and 3), primes, and the random rules (decreasing, increasing, and random numbers). The $\times C + K$ hypotheses cover any rule of the form $x_i = Cx_{i-1} + K$, 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. Since rules are based on the values of preceding numbers, we also need a scheme for generating the initial numbers in a sequence. We do this by sampling x_0 from a distribution assigning probability $1/|1 + x_0|$ to the positive and negative integers, and subsequent initial numbers from the same distribution centered around the preceding number. This acts as an implicit penalty against rules for which k_h is high, as they require more draws from this distribution.

The model defined in this section provides all we need to compute the posterior distribution over hypotheses given a sequence of numbers (Equation 2) and consequently to predict the next number in a sequence (Equation 1). Experiment 1 examines how well this model characterizes the predictions that people make about sequences of numbers.

Experiment 1: Predicting predictions

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 participants' and the model's ability to make judgments on simple and complex rules and when the given sequence was ambiguous as to the underlying rule. The stochastic pattern was chosen to demonstrate both participants and the model make sensible related judgments when the generating rule is not deterministic.

Methods

Participants A total of 146 undergraduates participated in the experiment for course credit or a free ice cream voucher.

Stimuli Five relational rules were tested: repeat the last number (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).

Procedure 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.

Results

As shown in Figure 1, the model and human prediction distributions are in close correspondence. The predictions shown for the model were obtained by optimizing the prior probability of the different hypothesis types to fit the human data, but are somewhat robust to variation in the prior. The correlation between the sets of predictions is $r = 0.87$. Since the increasing numbers pattern is 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 $r = 0.98$. The estimated prior probabilities of the seven types of hypotheses are: $\times C + K$ is 0.85, sum of the last two is 10^{-4}, mixtures of $\times C + K$ is 1.5×10^{-4}, repeat the last k_h numbers is 4.5×10^{-5}, i-th power is 0.05, primes is 6.9×10^{-7}, decreasing is 0.008, random is 0.006, and increasing is 0.09.

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. The analysis of optimal hypothesis testing given above predicts that people should seek to confirm the hypothesis that they assign highest posterior probability. To test this prediction, Experiment 2 manipulated the prior probability of different types of hypotheses to see if we could induce people to change which hypotheses they sought to confirm.

Experiment 2: Manipulating confirmation

Methods

Participants A total of 67 undergraduates 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.

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 target rule). Test selection was probed with 21 sequences consistent with both the sum of the last two numbers and the $\times C + K$ rule,

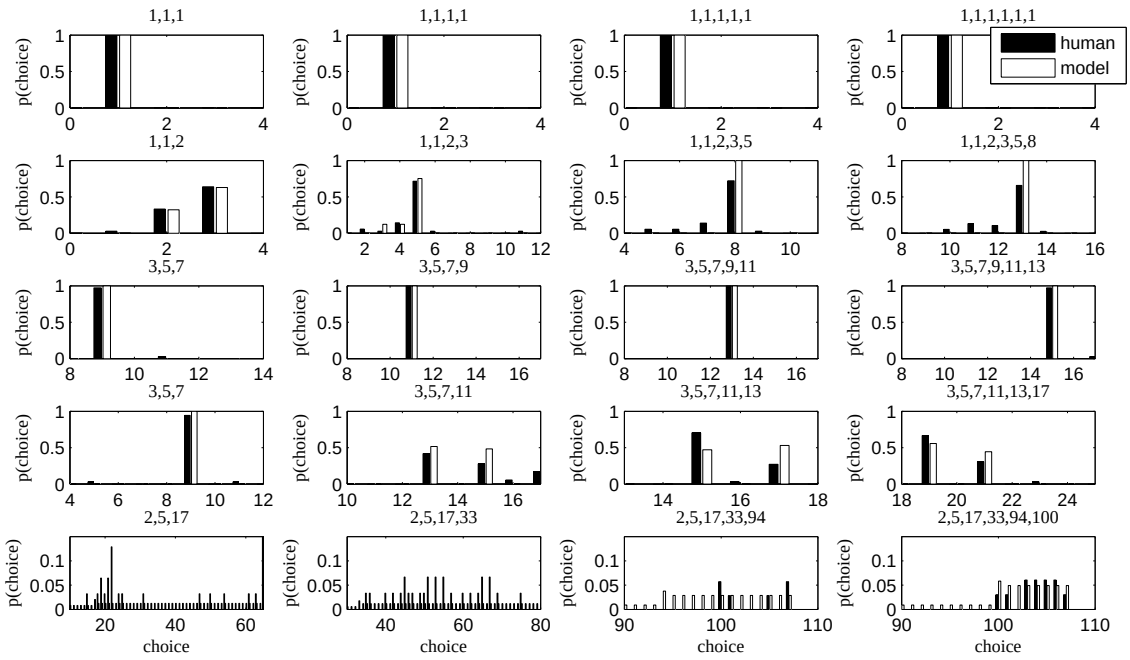

Figure 1: Results of Experiment 1. Each row of plots shows the predictions for one sequence as the number of elements increases from 3 to 7 across the columns. The five rules used to generate the sequences are (from top to bottom) repeating ones, sum of the last two numbers, increasing odd numbers, increasing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

shown to participants in all conditions. For example, one sequence, $(3,6,9)$, can be interpreted as $\times 1+3$ or the sum of the last two numbers $(3+6=9)$.

Procedure 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 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.

Results

If participants are sensitive to the prior probabilities of different environments, then they should choose to confirm the same rule as their training condition. Since 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 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, one blind to the hypothesis and both blind to condition, assigned the rules people selected as belonging to these three groups, with high inter-rater reliability ($\kappa = 0.90$). As the model predicts, participants were sensitive to the environment given in their training condition and changed their responses appropriately (see Figure 2). 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 significantly more often than participants in either the $\times C+K$ condition ($\chi^2(2) = 9.71, p < 0.01$) or the control condition ($\chi^2(2) = 196.25, p < 0.01$). Additionally, the responses for the sum of the last two numbers and control conditions were not significantly different ($\chi^2(2) = 1.11, p > 0.55$). 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.

Discussion and Conclusions

We have shown that the PTS is optimal under the assumption that the hypotheses under consideration are determinis-

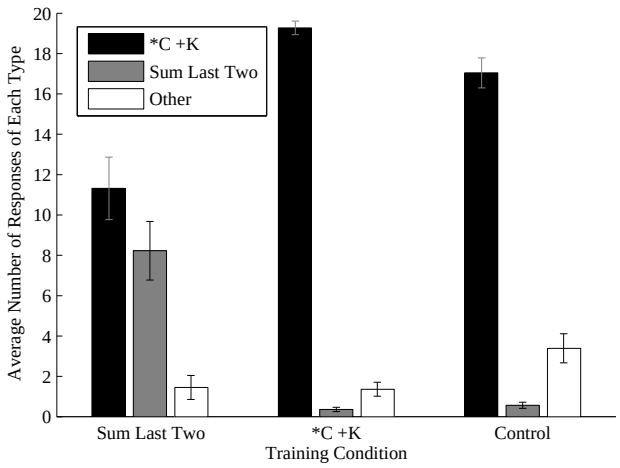

Figure 2: Results of Experiment 2, averaged over participants in each group. Error bars show one standard error.

tic, 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. In the first experiment, we showed that a Bayesian model of sequential prediction accurately characterizes human expectations. 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. In the second experiment, we demonstrated 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. We now consider how these results relate to previous work on the confirmation bias, and their implications for understanding why people might exhibit such a bias.

Klayman and Ha (1987) proposed exploring the role of the set of possible hypothesis on testing; however, few papers have used constrained hypothesis spaces to analyze hypothesis testing. One exception is Nelson and Movellan (2001) who explored directly applying the Bayesian generalization model and EIG to a task similar to the 2-4-6 task. 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 choices predicted by EIG matched the choices given by confirmation; however, in cases of low uncertainty, the choices predicted by EIG conflicted with choices given by confirmation (and human participants). One representive example where human responses deviate from EIG is for the given set $\{60,80,10,30\}$, the working hypothesis is multiples of ten, 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.

Our analysis indicates that the positive test strategy is optimal in a particular setting: when hypotheses are deterministic in their predictions. This is precisely the setting that people face in our numerical prediction task, where hypotheses are relational rules. However, in other settings – namely those where one hypothesis can be a superset of another – the PTS is suboptimal. In the spirit of previous rational analyses of confirmation (Oaksford & Chater, 1994), we propose explaining the fact that people pursue a suboptimal strategy in these non-deterministic settings as a consequence of assumptions about the structure of their environment – in our case, that rules are deterministic. If we live in a deterministic world, then choosing tests that confirm our expectations might be a simple adaptive strategy for this environment. We are in the process of developing a means of confirming this hypothesis.

Acknowledgments. We thank David McNamee and Kevin Canini for thoughtful discussions, our research assistants (especially Matt Cammann) for help running experiments, four anonymous reviewers for their comments, and the Air Force Office of Scientific Research (grant number FA9550-07-1-0351) and the UC Berkeley Chancellor's Faculty Partnership Fund for support.

References

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

Gelman, S. A., Coley, J. D., & Gottfried, G. M. (1994). Essentialist beliefs in children: The acquisition of concepts and theories. In *Mapping the mind: Domain specificity in cognition and culture* (p. 341-365).

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

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

Nelson, J. D., & Movellan, J. R. (2001). Active inference in concept learning. In *Advances in neural information processing systems* (Vol. 13, p. 45-51).

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

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

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

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

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.