In J. Schulkin (Ed.), *Action, perception and the brain: Adaptation and cephalic expression* (pp. 68-91). Basingstoke, United Kingdom: Palgrave Macmillan.

3

Homo Heuristicus and the Bias–Variance Dilemma

Henry Brighton and Gerd Gigerenzer

3

Homo Heuristicus and the Bias–Variance Dilemma

Henry Brighton and Gerd Gigerenzer

Introduction

Homo heuristicus makes inferences in uncertain environments using simple heuristics that ignore information (Gigerenzer and Brighton, 2009). Traditionally, heuristics are seen as second-best solutions which reduce effort at the expense of accuracy, and lead to systematic errors. The prevailing assumption is that, to understand the ability of humans and other animals to cope with uncertainty, one should investigate cognitive models that optimize. We introduced the term *Homo heuristicus* to highlight several reasons why this assumption can be misleading, and argue that heuristics play a critical role in explaining the ability of organisms to make accurate inferences from limited observations of an uncertain and potentially changing environment. In this chapter we use examples to sketch the theoretical basis for this assertion, and examine the progress made in the development of *Homo heuristicus* as a model of human decision-making.

What are heuristics and why study them?

Although frustratingly little detail is known about the mechanisms used by organisms to make inferences, illuminating insights into animal decision-making do exist. The ant species *Leptothorax albipennis* inhabit flat rock crevices, and select their nest site from a range of alternatives by estimating the relative area of potential nest using a heuristic. First, an ant will run around a potential nest on an irregular path for a fixed period of time while laying down a pheromone trail, and then leave. Later, the ant returns and runs around on a different irregular path, and then estimates the size of the site using the frequency with

which it reencounters the old trail. This heuristic is remarkably precise: nests half the area of others yielded reencounter frequencies 1.96 times greater (Mugford et al., 2001). Peahens also use a heuristic when choosing among potential mates. Rather than examining all the available peacocks, a peahen will investigate only three or four, and then choose the one with the largest number of eyespots (Petrie and Halliday, 1994). These two examples illustrate how organisms solve problems in an uncertain world using heuristics (see Hutchinson and Gigerenzer, 2005, for further examples).

What is a heuristic? The examples above highlight two hallmarks of heuristics: one-reason decision-making and limited search. Specifically, peahens could in principle integrate many features which describe potential mates, such as their size, dominance within the group, or volume of mating call. Instead, they use just one feature. Peahens could also exhaustively examine all the potential mates, but they instead consider just a handful. In the same way, ants tend to revisit potential nest sites a couple of times, rather than repeatedly approximating reencounter frequency. More generally, heuristics are best seen in contrast to optimization processes, which attempt to gather all available information, integrate this information, and derive predictions based on potentially complex models of the underlying processes which govern the observations. Heuristics, in contrast, are examples of what Herbert Simon (e.g., 1955; 1991) referred to as *satisficing* processes (a Northumbrian word for "satisfying"). Satisficing is the process of seeking a good-enough solution rather than seeking an optimal solution. For example, when selecting a good alternative from a series of options encountered sequentially, a satisficer might set an aspiration level, choose the first option that meets or exceeds this aspiration level, and then terminate search. Although heuristics tend to consume fewer processing resources as a result of ignoring information, this does not mean that they are less accurate than processes which consume more processing resources. Before considering these issues in greater detail, it is worth spelling out the relationship between optimal solutions, optimizing processes, and uncertainty.

Optimal solutions vs. the process of optimization

Consider a tin can manufacturer attempting to reduce costs by minimizing the surface area of the cans it produces. To package 12 ounces of soup, the manufacturer has calculated the height and width of the can which minimizes the amount of tin used. No other design uses less tin to package the same soup. This is an example of an optimal solution

to a problem: from the space of candidate solutions, the optimal solution is one which cannot be improved on. In this example, the relationship between the variables in question and the solution is certain. Solid geometry, coupled with the ability to take precise measurements, ensures a close fit between our model and the real world. The term "optimal" refers to such a state, but the term "optimization" refers to the process of searching for the optimal state. If we lacked knowledge of solid geometry, iteratively fine-tuning the tin can dimensions until the surface area is at a minimum would be a process of optimization. This process assumes that we can the measure the effect of our actions, and this measure serves as a proxy for performance. If we had knowledge of solid geometry, then we could derive the optimal solution directly. Both are examples of optimization.

Broadly speaking, optimization is any process which explicitly attempts to maximize some criterion assumed to be monotonically related to performance. As a consequence, optimization methods have a tendency to assume that more computation leads to greater precision. Optimization models are common in the study of cognition. Neural network models of cognition, for instance, usually rely on algorithms which attempt to minimize the error of the network by iteratively fine-tuning the strength of its synaptic weights (Rumelhart et al., 1986). While optimization is certainly a valid, widely practiced, and successful approach, it is not the only approach. Following Simon, we will use the term satisficing to refer to processes which do not optimize. Heuristics are precise specifications of these non-optimizing solutions (Gigerenzer et al., 1999).

Beyond optimization

Optimization methods are both widely studied and successful. Why, then, study heuristics? First of all, we may observe an organism satisficing, and focus on uncovering a precise understanding of the heuristic mechanism used by the organism. These findings lead naturally to the question of how ubiquitous satisficing mechanisms are in the natural world, and spur the search for underlying functional and information processing principles explaining why and how organisms might satisfice. For instance, the most common justification for heuristic processing is reduced resource consumption. Organisms consume resources such as time and energy when processing information, and, because these resources are limited, it is reasonable to assume a trade-off exists between accuracy and effort. Heuristics, according to this view, allow the decision-maker to make accurate inferences without expending the additional resources required to make the best inferences. This

justification is widely accepted, and rooted in the commonly assumed principle of the *accuracy–effort trade-off*:

> **Accuracy–effort trade-off**: information and computation cost time and effort; therefore minds rely on simple heuristics that are less accurate than strategies that use more information and computation.

Much of the following discussion will center on more fundamental reasons for examining heuristics. First, some problems are computationally intractable, and require, of necessity, decision-makers to satisfice. For instance, no human or machine can implement the optimal chess-playing strategy, due to its explosively large search space. For such problems, the only option may be to use heuristics. Second, another reason to look beyond optimization is the inherent uncertainty of natural environments. Organisms regularly face problems for which the optimal solution – the underlying data generating distribution, for example – is unknown or unknowable. Optimization in a world which is uncertain in this way is still possible, but less appealing, since we must knowingly optimize a misspecified criterion. Issues such as these open the door to alternative, potentially more appropriate, and superior information-processing models. This possibility leads to another reason for studying heuristics, which is to identify and explain less-is-more effects:

> **Less-is-more effects**: more information or computation can decrease accuracy; therefore, minds rely on simple heuristics in order to be more accurate than strategies that use more information and time.

The occurrence of less-is-more effects tells us that strategies which ignore information and limit search have the potential to better explain how organisms make accurate inferences in uncertain environments. Next, we will explore less-is-more effects, and work toward the study of the ecological rationality of heuristics, which examines in which environments a given strategy succeeds or fails, and why. These issues form part of a broader research program which aims at a systematic theory of heuristics that identifies their building blocks and the evolved capacities they exploit, and views the cognitive system as relying on an "adaptive toolbox" of heuristics.

Less-is-more effects

The term heuristic is often used with a negative connotation, suggesting a second-best solution to a problem better addressed by a more

principled, optimizing solution. The use of heuristics by people, according to this view, explains a number of human reasoning errors. The "Heuristics and biases" approach of Daniel Kahneman, Amos Tversky, and their collaborators emphasized that heuristics are sometimes good and sometimes bad, but placed a heavy focus on experiments designed to show that people often violate laws of logic, probability, or some other standard of rationality (Tversky and Kahneman, 1974). The association between heuristics and shoddy mental software is rooted in three widespread misconceptions:

1. Heuristics are always second-best.
2. We use heuristics only because of our cognitive limitations.
3. More time, more information, and more computation would always be better.

These three beliefs assume that the accuracy–effort trade-off, described above, holds. The inaccuracy of this overly simplistic picture can be demonstrated by considering perhaps the most widely used statistical model: the linear model fitted using least squares. Use of linear regression has become automatic among sociologists, economists, and psychologists when making inferences about their observations. Linear regression estimates the optimal beta weights for the predictors. In the 1970s, researchers discovered that unit weights (–1 or 1), or even random weights, can predict almost as accurately as, and sometimes better than, multiple linear regression (Dawes, 1979; Dawes and Corrigan, 1974; Einhorn and Hogarth, 1975; Schmidt, 1971). These less-is-more phenomena came as a surprise to the scientific community. When Robin Dawes presented the results at professional conferences, distinguished attendees told him that they were "impossible," his paper with Corrigan was first rejected and deemed "premature;" a sample of recent textbooks in econometrics revealed that none referred to the findings of Dawes and Corrigan (Hogarth, in press).

Before taking a closer look at less-is-more effects, what exactly does "more" refer to in the term "less-is-more?" For problems of inductive inference, the decision-maker's task is to process a series of observations with a view to identifying patterns of regularity among these observations. Identifying systematic regularities allows the decision-maker to make accurate predictions about novel or future observations. Consider, for instance, a weather forecaster observing past temperature trends in an attempt to spot a pattern which can be used to make better predictions. Problems like these highlight a crucial distinction

between two kinds of accuracy. First, *data fitting* is the ability to accurately describe already observed data. Experts find it alarmingly easy to explain, post hoc, why several hurricanes occurred in the previous year, or why the stock market crashed. *Data prediction* is much harder, and tells us to evaluate weather forecasters and financial experts by taking note of their past predictions, and checking how accurately they in fact predicted future events. Data prediction is the true test of a decision-maker's ability to make inductive inferences. In exactly the same way, an organism's inference mechanisms should help to accurately second-guess future events in its environment, rather than accurately describe past events. In short, when we refer to less-is-more effects in inductive inference, the "more" will refer to *predictive* accuracy, and the "less" will refer to various forms of ignoring information and limiting search.

Further less-is-more effects

Using predictive accuracy as the performance criterion, Czerlinski et al. (1999) conducted 20 studies comparing unit weighted regression (also known as tallying) and multiple regression. The models were compared using cross-validation, a process which repeatedly partitions the observations into one set used to estimate the model parameters, and another set used to measure the predictive accuracy of the fitted models (Stone, 1974). Specifically, Czerlinski et al. examined paired comparison tasks in which, for instance, the problem is to estimate which of two Chicago high schools will have a higher drop-out rate, based on cues such as writing score and proportion of Hispanic students. Ten of the 20 data sets were taken from a textbook on applied multiple regression (Weisberg, 1985). Averaged across all data sets, tallying achieved a higher predictive accuracy than multiple regression (Figure 3.1). Regression tended to overfit the data, as can be seen by the cross-over of lines: it had a higher fit than tallying, but a lower predictive accuracy.

These results illustrate that under certain circumstances tallying leads to higher predictive accuracy than multiple regression. They also illustrate that claims of universal superiority of one statistical model over another are rarely, if ever, true. Instead, they highlight the need to know in which environments simple tallying is more accurate than multiple regression, and in which environments it is not. This is the question of the *ecological rationality* of tallying. Early attempts to answer this question concluded that tallying succeeded when linear predictability of the criterion was moderate or small ($R^2 \leq 0.5$), the ratio of objects to cues was 10 or smaller, and the cues were correlated

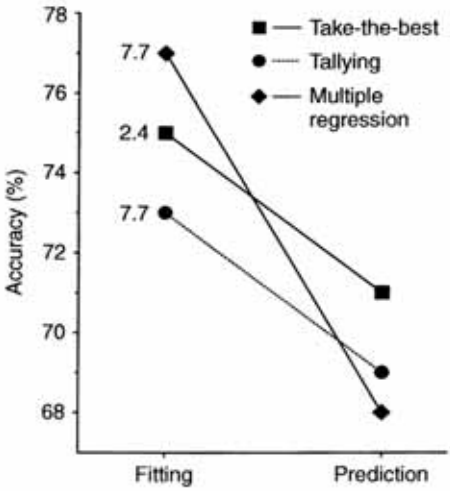

Figure 3.1 Less-is-more effects. Both tallying and take-the-best predict more accurately than multiple regression, despite using less information and computation. Note that multiple regression excels in data fitting ("hindsight"), that is, fitting its parameters to data that is already known, but performs relatively poorly in prediction ("foresight," as in cross-validation). Take-the-best is the most frugal, that is, it looks up, on average, only 2.4 cues when making inferences. In contrast, both multiple regression and tallying look up 7.7 cues on average. Here, we report mean accuracies taken with respect to 20 studies, including psychological, biological, sociological, and economic inference tasks (Czerlinski et al., 1999)

(Einhorn and Hogarth, 1975). The discovery that tallying can often match and even outperform complex calculations is important in understanding the nature of adaptive cognition. To what extent do findings such as these provide hints as to how the cognitive system makes inferences with the limited resources it has available? Note that the conditions under which tallying succeeds – low predictability of a criterion, small sample sizes relative to the number of available cues, and dependency between cues – is highly relevant for decision-making in natural environments.

Restricting attention to unit weights is one of many approaches to simplifying the basic linear model. Another simplification is to restrict attention to only a single cue when making decisions. The heuristic take-the-best uses this simplification when deciding between objects in a paired comparison task. First, take-the-best estimates the validity of each cue. Cue validity is the proportion of correct inferences the cue makes on previously observed comparisons. To make inferences

on novel, unobserved comparisons, take-the-best searches the cues in validity order sequentially until it finds a discriminating cue. A cue discriminates when the objects in question have different values for this cue, and therefore allow a decision to be made. Take-the-best always makes decisions on the basis of a single discriminating cue, although several cues may be examined before this cue is chosen. Figure 3.1 compares the performance of take-the-best with tallying and linear regression. Again, a less-is-more effect can be seen, but now the effect is more dramatic: take-the-best performs poorly in fitting the observations, but achieves a higher predictive accuracy than both tallying and linear regression.

Some researchers hypothesized that these less-is-more effects exploit the weakness of linear regression as a competitor. To address this point, we compared take-the-best with four additional inference strategies (Brighton, 2006). First, the *nearest neighbor classifier* is a nonlinear, widely used, and widely studied classification method which is known to perform well across many domains (Cover and Hart, 1967). Second, the tree induction algorithm *C4.5* is also widely studied and explicitly attempts to combat the problem of overfitting by pruning the decision trees it induces (Quinlan, 1993). Third, we compared take-the-best with another tree induction algorithm, *CART*, which tends to produce yet smaller trees than C4.5 (Breiman et al., 1994). Fourth, take-the-best will be compared with a slightly modified version of itself – *greedy take-the-best* – which commits the additional resources required to assess conditional dependencies between cues (Martignon and Hoffrage, 2002; Schmitt and Martignon, 2006). This variant of take-the-best orders cues by taking into account that cue validities, when estimated conditionally on the value of other cues, can differ from the validity estimates used by take-the-best. These estimates of conditional validity result in the induction of cue orders which differ from those selected by take-the-best.

Figure 3.2 compares the performance of take-the-best with these four alternative strategies in four environments taken from the Czerlinski et al. (1999) study. In all four environments, take-the-best outperforms the alternative mechanisms for most, if not all, sample sizes. Two points are worth making here. First, the less-is-more effects shown in Figure 3.1 and Figure 3.2 are to a certain degree robust, rather than hinging on a potentially "straw man" comparison with linear regression. Second, and more generally, it is statistically obvious that in some environments, and against some competitors, take-the-best (like any other model of inductive inference) will perform poorly.

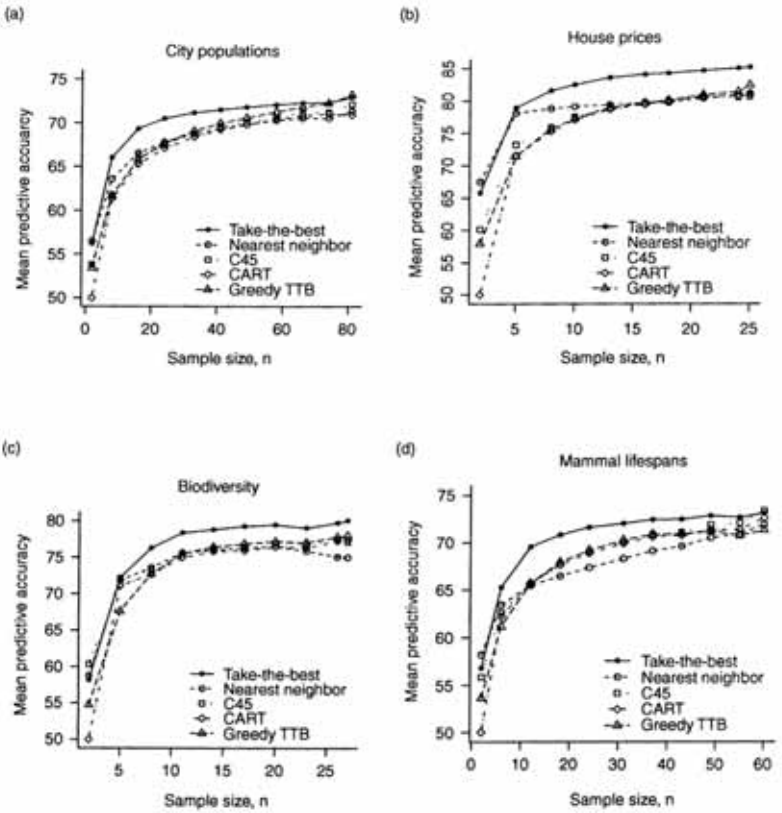

Figure 3.2 The performance of take-the-best in comparison to three well-known learning algorithms (nearest neighbor classifier, C4.5, and CART) and the greedy version of take-the-best, which orders cues by conditional validity. Mean predictive accuracy in cross-validation is plotted as a function of the size of the training sample for the task of deciding (a) which of two German cities has the larger population; (b) which of two houses has the higher price; (c) which of two Galapagos islands has greater biodiversity; and (d) which of two mammals is likely to live longer. These environments are taken from the study by Czerlinski et al. (1999)

As we noted above, the real question is to understand when and why certain forms of processing perform well, and when and why they perform poorly. Of particular interest here is the question of when and why ignoring information leads to superior performance, as we saw above.

Explaining less-is-more effects

Understanding *Homo heuristicus* requires understanding when and why less-is-more effects hold, and how an organism might exploit the existence of less-is-more effects as part of its design. The policies of simplifying weights and one-reason decision-making both reduce effort. They also increase accuracy, clearly demonstrating that the effort–accuracy trade-off is invalid as a general rule. Findings like these provide hints as to how organisms might make inferences in uncertain environments, where the task is to use limited observations to second-guess the predictive patterns underlying the observations. To understand the functioning of heuristics, we will again consider the task faced by a weather forecaster.

Over- and underfitting

The temperature in London on a given day of the year is uncertain but follows a seasonal pattern. Using the year 2000 as an example, we have plotted London's mean daily temperature in Figure 3.3(a). On top of these observations we have plotted two polynomial models that attempt to capture a systematic pattern in London's temperatures. The first model is a degree-3 polynomial (a cubic equation with four parameters), and the second is a degree-12 polynomial (which has 13 parameters). Comparing these two models, we see that the degree-12 polynomial captures monthly fluctuations in temperature while the degree-3 polynomial captures a simpler pattern charting a rise in temperature that peaks in the summer, followed by a slightly sharper fall.

If the weather forecaster is only interested in describing the past as accurately as possible, then picking the polynomial model that fits the data with the least error is the best option. This criterion would prefer the degree-12 polynomial. More generally, to maximize goodness of fit, higher and higher-degree polynomials could be chosen, with each added degree improving the ability of the model to capture minor fluctuations in temperature over smaller and smaller timescales. If London's daily temperatures for all subsequent years were guaranteed to match precisely those measured in the year 2000 then this approach would be ideal, because what we have observed in the past will continue to be observed in the future, and by describing the past more accurately, as with a higher-degree polynomial, we will also describe the future more accurately. The future is certain in this hypothetical world. As soon as uncertainty enters the picture, using goodness of fit to judge this ability

is a dangerous practice which will often lead to faulty conclusions (Pitt et al., 2002; Roberts and Pashler, 2000).

To estimate how well different models predict the future, based on observations of the past, we can reframe the task by estimating model parameters using only a sample of the observations and then test how well such models predict novel instances of the problem. More specifically, if we observe the temperature on 50 randomly selected days in the year 2000 and then fit a series of polynomial models of varying degree to this sample, we can measure how accurately each model goes on to predict the temperature on those days we did not observe in the year 2000. As a function of the degree of the polynomial model, the mean error in performing this prediction task is plotted in Figure 3.3(b). The model with the lowest mean error (with respect to many such samples of size 50) is a degree-4 polynomial – which shows that more complexity is not better. In short, Figure 3.3(b) tells us that the error in fitting the observations decreases as a function of the degree of the polynomial, which means that the best-predicting model would not have been chosen if we had judged models merely by checking how well they fit the observations. The most predictive model is very close to the lower bound of complexity, rather than at some intermediate or high level.

The bias–variance dilemma

Understanding how properties of a decision-maker's learning algorithm interact with properties of its task environment is a crucial step toward understanding how organisms can deal with uncertainty and error. To understand this problem we will adopt the perspective of an omniscient observer and consider the bias–variance dilemma (Geman et al., 1992; Hastie et al., 2001), a statistical perspective on the problem of inductive inference that decomposes prediction error into three components: a bias component, a variance component, and a noise component. Total prediction error is the sum of the following three terms:

$$\text{Error} = (\text{bias})^2 + \text{variance} + \text{noise}$$

This decomposition clarifies different sources of error, and how they are related to the properties of the learning algorithm. To illustrate this relationship, we will revisit the daily temperature example but change the rules of the game. The "true" underlying function behind London's mean daily temperatures is unknown. Nevertheless, we will put ourselves in the position of grand planner with full knowledge of the underlying function for the mean daily temperatures in some fictional

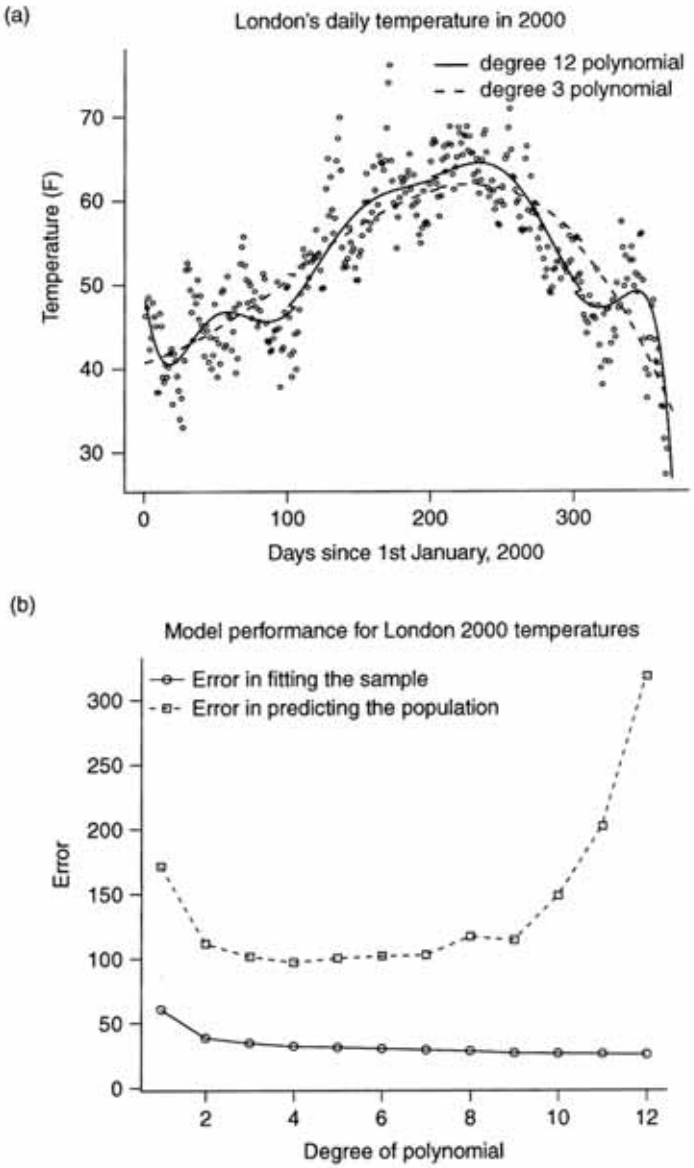

Figure 3.3 Plot (a) shows London's mean daily temperature in 2000, along with two polynomial models fitted using the least squares method. The first is a degree-3 polynomial, the second a degree-12 polynomial. Plot (b) shows both the mean error in fitting samples of 30 observations and the mean prediction error of the same models, both as a function of degree of polynomial

location. We denote this degree-3 polynomial function $h(x)$ and define it as

$$h(x) = 37 + \frac{15}{365}x + \frac{120}{365}x^2 + \frac{130}{365}x^3, \text{ where } 0 \leq x \leq 364.$$

Figure 3.4(a) plots this underlying trend for each day of the year. We will also assume that, when $h(x)$ is sampled, our observations suffer from normally distributed measurement error with $\mu = 0$ and $\sigma^2 = 4$. A random sample of 30 observations of $h(x)$ with this added error is shown on top of the underlying trend in Figure 3.4(a). If we now fit a degree-p polynomial to this sample of observations, and measure its error in approximating the function $h(x)$, can we draw a conclusion about the ability of degree-p polynomials to fit our "true" temperature function in general? Not really, because the sample we drew may be unrepresentative: it could result in a lucky application of our fitting procedure that perfectly models the underlying polynomial $h(x)$, or an unlucky one which results in high error. This single sample may not lead to a representative picture of the performance of degree-p polynomials in general, after other samples are taken into account.

A more reliable test of a model is to measure its mean accuracy by taking k random samples of size n, fitting a degree-p polynomial model to each one, and then considering this ensemble of models denoted by $y_1(x), y_2(x), \ldots, y_k(x)$. Figure 3.4(b) shows five polynomials of degree 2 resulting from $k = 5$ samples of $n = 30$ observations of $h(x)$. From the perspective of the organism, these samples can be likened to separate encounters with the environment, and the fitted polynomials likened to the responses of the organism to these encounters. The question now is how well a given type of model – here polynomials of degree 2 – captures the underlying function $h(x)$, which we can estimate by seeing how well the induced models perform on average, given their individual encounters with data samples. First consider the function $\bar{y}(x)$, which for each x gives the mean response of the ensemble of k polynomials:

$$\bar{y}(x) = \frac{1}{k}\sum y_i(x).$$

The *bias* of the model is the sum squared difference between this mean function and the true underlying function. Our omniscience is important now, because to measure the difference we need to know the underlying function $h(x)$. More precisely, bias is given by

$$(bias)^2 = \sum \left\{\bar{y}(x^n) - h(x^n)\right\}^2.$$

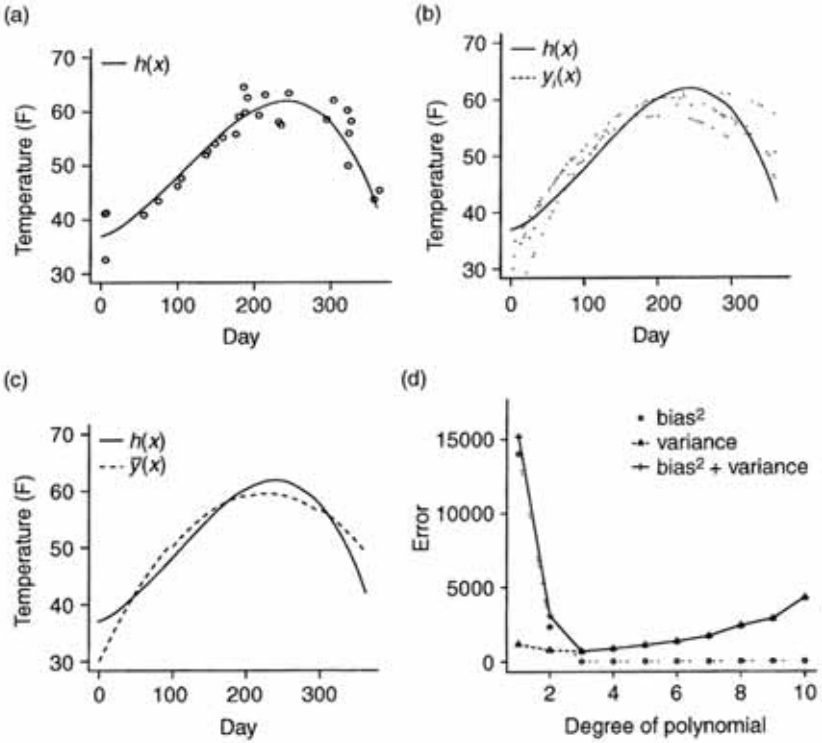

Figure 3.4 A fictional daily temperature function $h(x)$ used to illustrate bias and variance. (a) Graph of $h(x)$ and a sample of 30 points with added noise. (b) Five polynomials of degree-2, $y_i(x)$ for $1 \leq i \leq 5$, fitted to five further samples. (c) Mean of these five functions, $\bar{y}(x)$. Bias is the squared difference between $h(x)$ and $\bar{y}(x)$. Variance is the sum of the squared difference between each function $y_i(x)$ and $\bar{y}(x)$, measuring how much the induced functions vary about their mean. Plot (d) shows, as a function of degree of polynomial, the mean error in predicting the temperature on those days not in the observed samples, after fitting polynomials to samples of 30 noisy observations. This error is decomposed into bias and variance, also plotted as function of degree of polynomial

Figure 3.4(c) shows the $\bar{y}(x)$ arising from the five polynomials shown in figure 3.4(b). Assuming $k = 5$ is sufficient to provide us with a good estimate of $\bar{y}(x)$, this plot tells us that the model is biased, since it differs from $h(x)$. Zero bias is achieved if our average function is precisely the true function. Bias usually occurs when the model we use to explain the observations lacks the appropriate functional form to represent the true underlying function. In the absence of knowledge about the underlying function, bias can be reduced by making the space of models considered

by the learning algorithm sufficiently rich. However, this policy can simply replace one problem with another. Although the mean function response of the ensemble may capture the true underlying function without error, the individual models that contribute to this mean may each incur high error. That is, zero mean error can mask high error of the individual models. This source of error, which arises from the sensitivity of the learning algorithm to the contents of individual samples, is termed *variance*. Variance is the mean squared difference between each induced model function and the mean function:

$$\text{variance} = \sum_{n} \frac{1}{k} \sum \left\{ y_i(x^n) - \bar{y}(x^n) \right\}^2.$$

When variance increases as we consider more complex models, we say that these models are overfitting the data. The two properties of bias and variance reveal that the inductive inference of models involves a fundamental trade-off. We can aim to use a general purpose learning algorithm, such as a feed-forward neural network, that employs a wide and rich space of potential models, which more or less guarantees low bias. But, when we have a limited number of observations, the flexibility of the model space can incur a cost in high variance, since the learning algorithm is likely to induce a model that captures unsystematic variation. To combat high variance we can place restrictions on the model space and thereby limit the sensitivity of the learning algorithm to the vagaries of samples. But these restrictions run counter to the objective of general purpose inference, since they will necessarily cause an increase in bias for some problems.

This is the bias–variance dilemma. The balancing act required to achieve both low variance and low bias is clear in figure 3.4(d), which decomposes the error arising from polynomials from degree 1 (a straight line) to degree 10 at predicting our temperature function $h(x)$ from samples of size 30. For each polynomial degree we have plotted the bias (squared) of this type of model, its variance, and their sum. The polynomial degree that minimizes the total error is, not surprisingly, 3, because $h(x)$ is a degree-3 polynomial. Polynomial models of less than degree 3 suffer from bias, since they lack the ability to capture the underlying pattern. Polynomials of degree 3 or more have zero bias, as we would expect. But for polynomials of degree 4 or more the problem of overfitting arises, and their variance begins to increase due to their excess complexity. None of the models achieve zero error. This is due to the observation error we added when sampling, which corresponds to the noise term in the bias–variance decomposition.

Heuristics and the bias–variance dilemma

Bias and variance provide an insightful framework for examining the inference tasks faced by organisms. In particular, how do organisms control bias and variance, and keep them within acceptable limits? More specifically, how can the functioning of heuristics be understood in these terms? Recall Figure 3.1 and Figure 3.2, which showed how take-the-best outperformed several alternative inference strategies. Now we will analyze take-the-best by performing a bias–variance decomposition of its error. We will also decompose the error of its greedy counterpart, which we described above. As seen in Figure 3.2, the performance of the neural, exemplar, and decision tree models tend to be very similar to each other in paired comparison tasks, which in turn are very similar to the performance of the greedy version of take-the-best. Consequently, the performance of the greedy version of take-the-best provides a good proxy for the behavior of a number of alternative models of inductive inference.

Two artificially constructed environments will be used to compare the strategies. Both environments, therefore, will be governed by a known underlying functional relationship between the cues and criterion. Knowing these functional relationships will allow us to perform a bias–variance decomposition of the prediction error of the two strategies. The first environment is an instance of the class of *binary environments*, where the validity of the cues follows a noncompensatory pattern, and all cues are uncorrelated. An environment has a noncompensatory pattern when the validity of the cues decays rapidly as a function of their rank in the cue order. Noncompensatory environments are one example of a class of environments for which we have analytic results showing that take-the-best is unbiased and likely to perform well (Katsikopoulos and Martignon, 2006; Martignon and Hoffrage, 2002). The second environment used in our comparison, however, is an instance of the class of *Guttman environments*, inspired by the Guttman scale (Guttman, 1944), in which all the cues are maximally correlated with criterion, and all have a maximum validity of 1. Formal definitions and illustrations of both these environments are provided in Appendix 1 of Gigerenzer and Brighton (2009).

Figure 3.5(a–d) plots, for both of these environments, the prediction error achieved by take-the-best and its greedy counterpart. The performance of each model is shown separately in order to clearly distinguish the bias and variance components of error, which, when added together, comprise the total prediction error. Three findings are

revealed. First, in the binary environment, take-the-best performs worse than its greedy counterpart. This result illustrates that analytic results detailing when take-the-best is unbiased will not necessarily help to explain when take-the-best performs well. Second, in the Guttman environment, take-the-best outperforms its greedy counterpart. This result illustrates that proving that another strategy achieves a better fit than take-the-best is something quite different from proving that the strategy also achieves a higher predictive accuracy. Third, and perhaps most importantly, Figure 3.5 reveals that both of these behaviors are driven by the variance component of error, and the relative ability of the two strategies to keep variance within acceptable limits. Bias plays almost no role in explaining the difference in performance between the models, and the less-is-more effect we demonstrated in Figure 3.2 can also be explained by the relative ability of the models to control variance. In short, this comparison tells us that take-the-best bets on the fact that ignoring dependencies between cues is likely to result in low variance. Model comparisons in natural environments show that this bet is often a good one. But, as this comparison has revealed, the bet can also fail, even when take-the-best has zero bias.

At this point, it is important to note that the concepts of bias and variance have allowed us to move beyond simply labeling the behavior of an induction algorithm as "overfitting the data," or "suffering from excess complexity," because the relative ability of two algorithms to avoid these pathologies will always depend on the amount of data available. First of all, from the perspective of bias, take-the-best offers no advantage over the alternative methods we have considered, because practically all models of inductive inference are capable of capturing the same systematic patterns in data as take-the-best. Consequently, if a heuristic like take-the-best is to outperform an alternative method, it must do so by incurring less variance. Second, the variance component of error is always an interaction between characteristics of the inference strategy, the structure of the environment, and the number of observations available. Thus, saying that a heuristic works because it avoids overfitting the data is only a shorthand explanation for what is often a more complex interaction between the heuristic, the environment, and the sample size.

Bias, variance, and cognition

Organisms experience a limited number of observations, and cannot be expected to know, or possess the ability to model, underlying environmental regularities without error. The former constraint tells us that an organism's inference mechanisms must control the variance

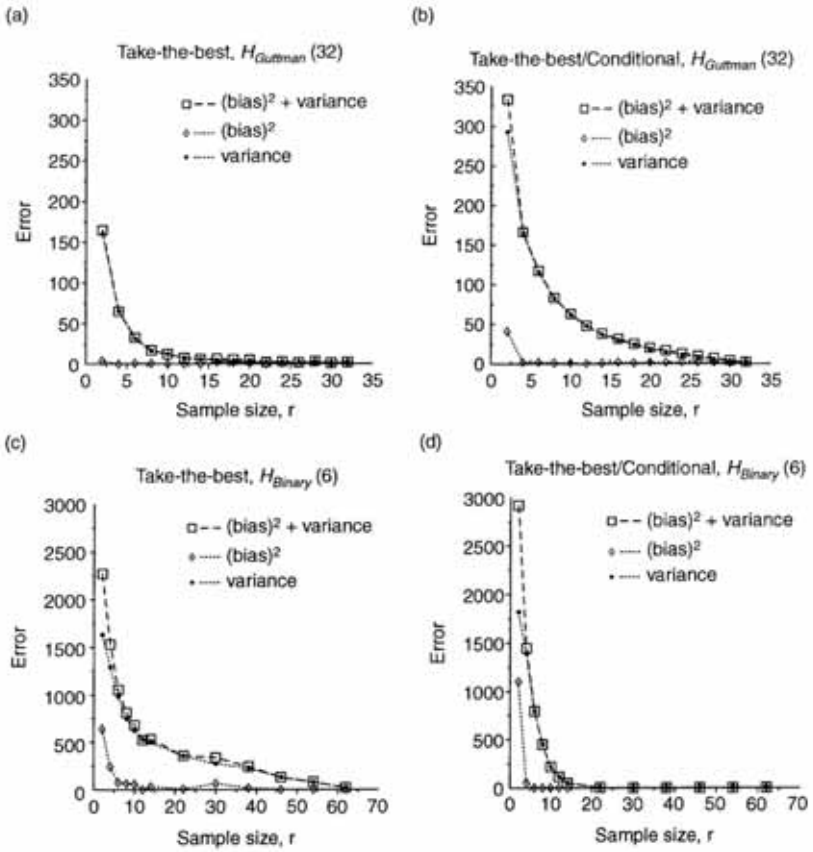

Figure 3.5 An illustration of the role played by variance in the performance of take-the-best. Plots (a) and (b) illustrate that, in Guttman environments, take-the-best outperforms the greedy variant of this heuristic, which orders cues by conditional validity. The performance difference is due to variance alone. Plots (c) and (d) illustrate that variance also explains why take-the-best is outperformed in binary environments. In both cases, take-the-best is unbiased and the relative performance of the models is explained almost entirely by variance

component of error in order to make accurate inferences, while the latter constraint tells us that bias is inevitable. Taken together, considerations of bias and variance suggest that the apparent ability of humans and other animals to generalize accurately from limited observations of an uncertain environment is likely to hinge on the use of biased inference mechanisms which excel at limiting variance. Our analysis of

take-the-best has shown that heuristics work for precisely this reason. The other models considered opt to perform some form of optimization over a rich collection of potential hypotheses. Yet take-the-best can make more accurate inferences by being specialized (and therefore biased), and limiting variance by ignoring information.

This observation is the basis for our assertion that "biased minds make better inferences" (Gigerenzer and Brighton, 2009, p. 107). More generally, several established statistical models use misspecified models in order to reduce variance. Ridge regression is one well-known example, and works by limiting, or "squashing," the fitted parameters' values in a linear model (Hoerl and Kennard, 2000). By studying less-is-more effects, we are essentially exploring the question of how ignoring information and limiting the use of computation resources can introduce bias, but offset this bias with a greater reduction in variance. Heuristics are one way of exploring this question. The study of ecological rationality asks in which environments these heuristic tricks work, and in which will they fail. The notion of *Homo heuristicus* asserts that the cognitive system relies on such tricks in order to make accurate inferences in uncertain environments, while at the same time using limited processing resources.

Homo Heuristicus relies on an adaptive toolbox

Let us return to our original proposition, that *Homo heuristicus* accurately captures fundamental aspects of human decision-making. We set the scene for this claim using examples of less-is-more effects, and explained them using the bias–variance dilemma. Throughout this discussion we have contrasted heuristics with optimization processes, and argued that, performing by ignoring information, limiting search, and relying on potentially impoverished representations of the world, an organism can make accurate inferences from limited observations. In our research group, this functional approach has been developed in parallel to a significant body of empirical work investigating human use of heuristics. Although this discussion has centered on functional concerns, both approaches are essential to developing a more detailed understanding of what we term the *adaptive toolbox*, which is a metaphor used to conceptualize the stock processing strategies available to the organism.

Table 3.1 lists 10 heuristics we view as being in the adaptive toolbox of humans, along with some of the surprising findings they have led to. Some of these heuristics address the same task, but perform well in different kinds of environment. For example, tallying and take-the-best

Table 3.1 Ten well-studied heuristics for which there is evidence that they are in the adaptive toolbox of humans. Each heuristic can be used to solve problems in social and nonsocial environments. See the references given for more information regarding their ecological rationality, and the surprising predictions they entail

Heuristic	Definition[1]	Ecologically rational if:	Surprising findings (examples)
Recognition heuristic (Goldstein and Gigerenzer, 2002)	If one of two alternatives is recognized, infer that it has the higher value on the criterion.	Recognition validity >0.5	Less-is-more effect if $\alpha>\beta$; systematic forgetting can be beneficial (Schooler and Hertwig, 2005).
Fluency heuristic (Jacoby and Dallas, 1981)	If both alternatives are recognized but one is recognized faster, infer that it has the higher value on the criterion.	Fluency validity >0.5	Less-is-more effect; systematic forgetting can be beneficial (Schooler and Hertwig, 2005)
Take-the-best (Gigerenzer and Goldstein, 1996)	To infer which of two alternatives has the higher value: (1) search through cues in order of validity, (2) stop search as soon as a cue discriminates, and (3) choose the alternative this cue favors.	Cue validities are heavily skewed (Katsikopoulos and Martignon, 2006; Martignon and Hoffrage, 2002).	Often predicts more accurately than multiple regression (Czerlinski et al. 1999); neural networks, exemplar models, and decision tree algorithms (Brighton, 2006).
Tallying (unit-weight linear model, Dawes, 1979)	To estimate a criterion, do not estimate weights but simply count the number of positive cues.	Cue validities vary little, low redundancy (Hogarth and Karelaia, 2005; 2006).	Often predict with equal or greater accuracy than multiple regression (Czerlinski et al., 1999).
Satisficing (Simon, 1955; Todd and Miller, 1999)	Search through alternatives and choose the first one that exceeds your aspiration level.	Number of alternatives decreases rapidly over time, such as in seasonal mating pools (Dudey and Todd, 2002).	Aspiration levels can lead to significantly better choices than chance, even if they are arbitrary (e.g., the secretary problem, see Gilbert and Mosteller, 1966; the envelope problem, see Bruss, 2000).

Continued

Table 3.1 Continued

Heuristic	Definition[1]	Ecologically rational if:	Surprising findings (examples)
1/*N*; equality heuristic (DeMiguel et al., 2009)	Allocate resources equally to each of *N* alternatives.	High unpredictability, small learning sample, large *N*.	Can outperform optimal asset allocation portfolios.
Default heuristic (Johnson and Goldstein, 2003)	If there is a default, do nothing.	Values of those who set defaults match those of the decision-maker, when the consequences of a choice are hard to foresee.	Explains why mass mailing has little effect on organ donor registration; predicts behavior when trait and preference theories fail.
Tit-for-tat (Axelrod, 1984)	Cooperate first and then imitate your partner's last behavior.	The other players also play tit-for-tat; the rules of the game allow defection or cooperation but not divorce.	Can lead to a higher payoff than optimization (backward induction).
Imitate the majority (Boyd and Richerson, 2005)	Consider the majority of people in your peer group and imitate their behavior.	Environment is stable or only changes slowly; info search is costly or time-consuming.	A driving force in bonding, group identification, and moral behavior.
Imitate the successful (Boyd and Richerson, 2005)	Consider the most successful person and imitate his or her behavior.	Individual learning is slow; information search is costly or time-consuming.	A driving force in cultural evolution.

[1]For formal definitions, see references.

both describe how decision-makers make paired comparisons, but their predictive relative accuracy will depend on the statistical properties of the environment. Other heuristics address different tasks, but share common design features. For example, an organism relying on the default heuristic makes no decision if the environment already supplies a default option. An organism relying on the recognition heuristic opts for a recognized alternative when the other is unrecognized. Both heuristics implement one-reason decision-making, which means that they focus on one source of information, rather than integrating several potentially relevant additional sources of information, when making a decision. Notice, also, that none of the heuristics in Table 3.1

implement the process of optimization. For a more thorough discussion of the heuristics in Table 3.1, we refer the reader to Gigerenzer and Brighton (2009).

Several fundamental questions remain. For example, how do organisms select between strategies in the adaptive toolbox, and how can we understand, more generally, the functional relationship between cognitive mechanisms and the structure of the environment? These two questions are closely related, and center on the basic question of understanding the adaptive relationship between an organism and its environment. Such issues are clearly not specific to the study of heuristics, but the study of heuristics has proven a productive way of exploring them. Faced with the problem of designing an organism capable of functioning in a certain environment, it would be functional to equip the organism with a rich and accurate model of its environment, such that the consequences of its actions could be predicted. Such certainty allows the organism "to look before you leap." In highly uncertain environments, one must face the unavoidable conclusion that accurate models of the world are beyond reach, observations will be limited, and error is inevitable. The study of heuristics is the study of simple strategies which respond to this problem, and the notion of *Homo heuristicus* proposes that these heuristics play a fundamental role in how the cognitive systems of humans and other animals respond so successfully to environmental uncertainty.

References

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Boyd, R., and Richerson, P. J. (2005). *The origin and evolution of cultures*. New York: Oxford University Press.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, P. J. (1994). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Brighton, H. (2006). Robust inference with simple cognitive models. In C. Lebiere, and R. Wray (Eds), *AAAI Spring Symposium: Cognitive Science Principles Meet AI-Hard Problems*. Menlo Park, CA: American Association for Artificial Intelligence, pp. 17–22.

Bruss, F. T. (2000). Der Ungewissheit ein Schnippchen schlagen. *Spektrum der Wissenschaft, 6*, 106.

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*, 21–7.

Czerlinski, J., Gigerenzer, G., and Goldstein, D. G. (1999). How good are simple heuristics? In G. Gigerenzer, P. M. Todd, and the ABC Research Group, *Simple heuristics that make us smart*. New York: Oxford University Press, pp. 97–118.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–82.

Dawes, R. M., and Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95–106.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies, 22,* 1, 915–53.

Dudey, T., and Todd, P. M. (2002). Making good decisions with minimal information: Simultaneous and sequential choice. *Journal of Bioeconomics, 3,* 195–215.

Einhorn, H. J., and Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13,* 171–92.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4,* 1–58.

Gigerenzer, G., and Brighton, H. (2009). *Homo heuristicus*: Why biased minds make better inferences. *Topics in Cognitive Science, 1,* 107–43.

Gigerenzer, G., and Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–69.

Gigerenzer, G., Todd, P. M., and the ABC Research Group (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Gilbert, J. P., and Mosteller, F. (1966). Recognizing the maximum of a sequence. *American Statistical Association Journal, 61,* 35–73.

Goldstein, D. G., and Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109,* 75–90.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139–50.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer.

Hoerl, A. E., and Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 42,* 80–6.

Hogarth, R. M. (in press). When simple is hard to accept. In P. M. Todd, G. Gigerenzer, and the ABC Research Group, *Ecological rationality: Intelligence in the world.* Oxford, UK: Oxford University Press.

Hogarth, R. M., and Karelaia, N. (2005). Ignoring information in binary choice with continuous variables: When is less "more"? *Journal of Mathematical Psychology, 49,* 115–24.

Hogarth, R. M., and Karelaia, N. (2006). "Take-the-best" and other simple strategies: Why and when they work "well" with binary cues. *Theory and Decision, 61,* 205–49.

Hutchinson, J. M. C., and Gigerenzer, G. (2005). Simple heuristics and rules of thumb: Where psychologists and behavioural biologists might meet. *Behavioural Processes, 69,* 97–124.

Katsikopoulos, K. V., and Martignon, L. (2006). Naive heuristics for paired comparisons: Some results on their relative accuracy. *Journal of Mathematical Psychology, 50,* 488–94.

Jacoby, L. L., and Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110,* 306–40.

Johnson, E. J., and Goldstein, D. G. (2003). Do defaults save lives? *Science, 302,* 1,338–9.

Martignon, L., and Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparisons. *Theory and Decision, 52,* 29–71.

Mugford, S. T., Mallon, E. B., and Franks, N. R. (2001). The accuracy of Buffon's needle: A rule of thumb used by ants to estimate area. *Behavioral Ecology, 12,* 655–8.

Petrie, M., and Halliday, T. (1994). Experimental and natural changes in the peacock's (Pavo cristatus) train can affect mating success. *Behavioral Ecology and Sociobiology, 35,* 213–7.

Pitt, M. A., Myung, I. J., and Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–91.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Roberts, S., and Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107,* 358–67.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, and J. L. McClelland (Eds), *Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 1: Foundations).* Cambridge, MA: MIT Press, pp. 318–62.

Schmidt, F. L. (1971). The relative efficiency of regression and simple unit weighting predictor weights in applied differential psychology. *Educational and Psychological Measurement, 31,* 699–714.

Schmitt, M., and Martignon, L. (2006). On the complexity of learning lexicographic strategies. *Journal of Machine Learning Research, 7,* 55–83.

Schooler, L. J., and Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112,* 610–28.

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

Simon, H. A. (1991). *Models of my life.* New York: Basic Books.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B, 36,* 111–47.

Todd, P. M., and Miller, G. F. (1999). From pride and prejudice to persuasion: Realistic heuristics for mate search. In G. Gigerenzer, P. M. Todd, and the ABC Research Group, *Simple heuristics that make us smart.* New York: Oxford University Press, pp. 287–308.

Tversky, A., and Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185,* 1,124–31.

Weisberg, S. (1985). *Applied linear regression.* New York: Wiley.