Appetizer

CHAPTER 1

Homo heuristicus: Why Biased Minds Make Better Inferences

Gerd Gigerenzer and Henry Brighton

Abstract: Heuristics are efficient cognitive processes that ignore information. In contrast to the widely held view that less processing reduces accuracy, the study of heuristics shows that less information, computation, and time can in fact improve accuracy. We review the major progress made so far: (a) the discovery of less-is-more effects; (b) the study of the ecological rationality of heuristics, which examines in which environments a given strategy succeeds or fails, and why; (c) an advancement from vague labels to computational models of heuristics; (d) the development of 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;" and (e) the development of an empirical methodology that accounts for individual differences, conducts competitive tests, and has provided evidence for people's adaptive use of heuristics. Homo heuristicus has a biased mind and ignores part of the available information, yet a biased mind can handle uncertainty more efficiently and robustly than an unbiased mind relying on more resource-intensive and general-purpose processing strategies.

As far as we can know, animals have always relied on heuristics to solve adaptive problems, and so have humans. To measure the area of a candidate nest cavity, a narrow crack in a rock, an ant has no yardstick but a rule of thumb: Run around on an irregular path for a fixed period while laying down a pheromone trail, and then leave. Return, move around on a different irregular path, and estimate the size of the cavity by the frequency of encountering the old trail. This heuristic is remarkably precise: Nests half the area of others yielded reencounter frequencies 1.96 times greater (Mugford, Mallon, & Franks, 2001). To choose a mate, a peahen similarly uses a heuristic: Rather than investigating all peacocks posing and displaying in a lek eager to get her attention or weighting and adding all male features to calculate the one with the highest expected utility, she investigates only three or four, and chooses the one with the largest number of eyespots (Petrie & Halliday, 1994). Many of these evolved rules of thumb are amazingly simple and efficient (for an overview, see Hutchinson & Gigerenzer, 2005).

The Old Testament says that God created humans in his image and let them dominate all animals, from whom they fundamentally differ (Genesis 1:26). It might not be entirely accidental that in cognitive science some form of omniscience (knowledge of all relevant probabilities and utilities, for instance) and omnipotence (the ability to compute complex functions in a split second) has shaped models of human cognition. Yet humans and animals have common ancestors, related sensory and motor processes, and even share common cognitive heuristics.

Consider how a baseball outfielder catches a ball. The view of cognition favoring omniscience and omnipotence suggests that complex problems are solved with complex mental algorithms: The player "behaves as if he had solved a set of differential equations in predicting the trajectory of the ball … At some subconscious level, something functionally equivalent to the mathematical calculations is going on" (Dawkins, 1989, p. 96). Dawkins carefully inserts "as if" to indicate that he is not quite sure whether brains actually perform these computations. And there is indeed no evidence that brains do. Instead, experiments have shown that players rely on several heuristics. The gaze heuristic is the simplest one and works if the ball is already high up in the air: Fix your gaze on the ball, start running, and adjust your running speed so that the angle of gaze remains constant (see Gigerenzer, 2007). A player who relies on the gaze heuristic can ignore all causal variables necessary to compute the trajectory of the ball—the initial distance, velocity, angle, air resistance, speed and direction of wind, and spin, among others. By paying attention to only one variable, the player will end up where the ball comes down without computing the exact spot. The same heuristic is also used by animal species for catching prey and for intercepting potential mates. In pursuit and predation, bats, birds, and dragonflies maintain a constant optical angle between themselves and their prey, as do dogs when catching a Frisbee (Shaffer, Krauchunas, Eddy, & McBeath, 2004).

The term *heuristic* is of Greek origin, meaning "serving to find out or discover." The mathematician George Polya distinguished heuristics from analytic methods; for instance, heuristics are indispensable for finding a proof, whereas analysis is required to check a proof's validity. In the 1950s, Herbert Simon (1955, 1991), whose collaborator Allen Newell studied with Polya, first proposed that people satisfice rather than maximize. *Maximization* means optimization, the process of finding the best solution for a problem, whereas *satisficing* (a Northumbrian word for "satisfying") means finding a good-enough solution. Simon used his term *satisficing* both as a generic term for everything that is not optimizing as well as for a specific heuristic: In order to select a good alternative (e.g., a house or a spouse) from a series of options encountered sequentially, a person sets an aspiration level, chooses the first one that meets the aspiration, and then terminates search. The aspiration level can be fixed or adjusted following experience (Selten, 2001). For Simon, humans rely on heuristics not simply because their cognitive limitations prevent them from optimizing but also because of the task environment. For instance, chess has an optimal solution, but no computer or mind, be it Deep Blue or Kasparov, can find this optimal sequence of moves, because the sequence is computationally intractable to discover and verify. Most problems of interest are computationally intractable, and this is why engineers and artificial intelligence (AI) researchers often rely on heuristics to make computers smart.

In the 1970s, the term *heuristic* acquired a different connotation, undergoing a shift from being regarded as a method that makes computers smart to one that explains why people are not smart. Daniel Kahneman, Amos Tversky, and their collaborators published a series of experiments in which people's reasoning was interpreted as exhibiting fallacies. "Heuristics and biases" became one phrase. It was repeatedly emphasized that heuristics are sometimes good and sometimes bad, but virtually every experiment was designed to show that people violate a law of logic, probability, or some other standard of rationality. On the positive side, this influential research drew psychologists' attention to cognitive heuristics and helped to create two new fields: behavioral economics, and behavioral law and economics. On the negative side, heuristics became seen as something best to avoid, and consequently, this research was disconnected from the study of heuristics in AI and behavioral biology. Another negative and substantial consequence was that computational models of heuristics, such as lexicographic rules (Fishburn, 1974) and elimination-by-aspects (Tversky, 1972), became replaced by one-word labels: availability, representativeness,

and anchoring. These were seen as the mind's substitutes for rational cognitive procedures. By the end of the 20th century, the use of heuristics became associated with shoddy mental software, generating three widespread misconceptions:

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

These three beliefs are based on the so-called *accuracy-effort trade-off*, which is considered a general law of cognition: If you invest less effort, the cost is lower accuracy. Effort refers to searching for more information, performing more computation, or taking more time; in fact, these typically go together. Heuristics allow for fast and frugal decisions; thus, it is commonly assumed that they are second-best approximations of more complex "optimal" computations and serve the purpose of trading off accuracy for effort. If information were free and humans had eternal time, so the argument goes, more information and computation would always be better. For instance, Tversky (1972, p. 98) concluded that elimination-by-aspects "cannot be defended as a rational procedure of choice." More outspokenly, two eminent decision theorists, Keeney and Raiffa (1993), asserted that reliance on lexicographic heuristics "is more widely adopted in practice than it deserves to be," and they stressed that "it is our belief that, leaving aside 'administrative ease,' it is rarely appropriate" and "will rarely pass a test of reasonableness" (pp. 77–78). They did not, however, put their intuition to a test. A few years later, our research team conducted such tests, with surprising results. Contrary to the belief in a general accuracy-effort trade-off, less information and computation can actually lead to higher accuracy, and in these situations the mind does not need to make trade-offs. Here, a *less-is-more* effect holds.

That simple heuristics can be more accurate than complex procedures is one of the major discoveries of the last decades (Gigerenzer, 2008). Heuristics achieve this accuracy by successfully exploiting evolved mental abilities and environmental structures. Since this initial finding, a systematic science of heuristics has emerged, which we review in this article, together with the reactions it has provoked. Beginning with the standard explanation of why people rely on heuristics and the common assumption of a general accuracy-effort trade-off, we introduce less-is-more phenomena that contradict it. We then present the ecological rationality of heuristics as an alternative explanation and show how less-is-more phenomena emerge from the bias–variance dilemma that cognitive systems face in uncertain worlds. In the following sections, we make a case against the widespread use of vague labels instead of models of heuristics, review some of the progress made in the development of a systematic theory of the adaptive toolbox, and end with a discussion of proper methodology.

1. THE DISCOVERY OF LESS-IS-MORE

More information is always better. This has been argued both explicitly and implicitly. The philosopher Rudolf Carnap (1947) proposed the "principle of total evidence," the recommendation to use all the available evidence when estimating a probability. The statistician I. J. Good (1967) argued, similarly, that it is irrational to leave observations in the record but not use them. In the same way, many theories of cognition—from exemplar models to prospect theory to Bayesian models of cognition—assume that all pieces of information should be combined in the final judgment. The classical critique of these models is that in the real world, search for information costs time or money, so there is a point where the costs of further search are no longer justified. This has led to optimization-under-constraints theories in which search in the world (e.g., Stigler, 1961) or in memory (e.g., Anderson, 1991) is terminated when the expected costs exceed the benefits. Note that in this "rational analysis of cognition," more information is still considered better, apart from its costs. Similarly, the seminal analysis of the adaptive decision maker (Payne, Bettman, &

Johnson, 1993) rests on the assumption that the rationale for heuristics is a trade-off between accuracy and effort, where effort is a function of the amount of information and computation consumed:

> *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.

Here is the first important discovery: Heuristics can lead to more accurate inferences than strategies that use more information and computation (see below). Thus, the accuracy-effort trade-off does not generally hold; there are situations where one attains higher accuracy with less effort. Even when information and computation are entirely free, there is typically a point where less is more:

> *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.

To justify the use of heuristics by accuracy-effort trade-offs means that it is not worth the effort to rely on more complex estimations and computations. A less-is-more effect, however, means that minds would not gain anything from relying on complex strategies, even if direct costs and opportunity costs were zero. Accuracy-effort trade-offs are the conventional justification for why the cognitive system relies on heuristics (Beach & Mitchell, 1978; Shah & Oppenheimer, 2008), which refrains from any normative implications. Less-is-more effects are a second justification with normative consequences: They challenge the classical definition of rational decision making as the process of weighting and adding all information. Note that the term *less-is-more* does not mean that the less information one uses, the better the performance. Rather, it refers to the existence of a point at which more information or computation becomes detrimental, independent of costs. In this article, when we refer to less information, we refer to ignoring cues, weights, and dependencies between cues. Our discussion of less-is-more effects begins with the seminal work of Robin Dawes and colleagues on ignoring weights.

1.1. Ignoring Weights to Make Better Predictions: Tallying

From sociologists to economists to psychologists, social scientists routinely rely on multiple linear regression to understand social inequality, the market, and individual behavior. Linear regression estimates the optimal beta weights for the predictors. In the 1970s, researchers discovered that equal (or random) weights can predict almost as accurately as, and sometimes better than, multiple linear regression (Dawes, 1979; Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975; Schmidt, 1971). Weighting equally is also termed *tallying*, reminiscent of the tally sticks for counting, which can be traced back some 30,000 years in human history. These results 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," and a sample of recent textbooks in econometrics revealed that none referred to the findings of Dawes and Corrigan (Hogarth, in press).

In these original demonstrations, there was a slight imbalance: Multiple regression was tested by cross-validation (that is, the model was fitted to one half of the data and tested on the other half) but tallying was not. Czerlinski, Gigerenzer, and Goldstein (1999) conducted 20 studies in which both tallying and multiple regression were tested by cross-validation, correcting for this imbalance. All tasks were paired comparisons; for instance, estimating which of two Chicago high schools will have a higher dropout 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 1-1). 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.

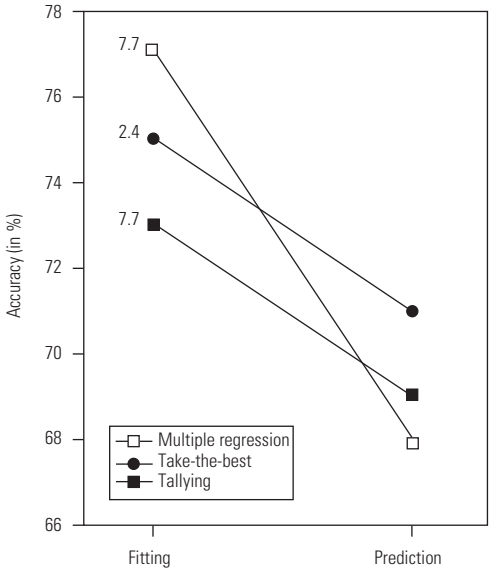

Figure 1-1. Less-is-more effects. Both tallying and take-the-best predict more accurately than multiple regression, despite using less information and computation. Note that multiple regression excels in data fitting ("hindsight"), that is, fitting its parameters to data that are already known, but performs 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. The results shown are averaged across 20 studies, including psychological, biological, sociological, and economic inference tasks (Czerlinski, Gigerenzer, & Goldstein, 1999). For each of the 20 studies and each of the three strategies, the 95% confidence intervals were ≤.4 percentage points.

The point here is not that tallying leads to more accurate predictions than multiple regression. The real and new question is 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 indicated that tallying succeeded when linear predictability of the criterion was moderate or small ($R^2 \leq .5$), the ratio of objects to cues was 10 or smaller, and the cues were correlated (Einhorn & Hogarth, 1975). The discovery that tallying can often match and even outperform complex calculations is relevant to understanding the nature of adaptive cognition. Why should a mind waste time and effort in estimating the optimal weights of cues if they do not matter or even detract from performance? 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—are not infrequent in natural environments. Yet many models of cognition continue to assume that the weighting of cues is a fundamental characteristic of cognitive processing. Why is this discovery of a less-is-more effect neglected? According to the tools-to-theories heuristic (Gigerenzer, 1991), models of cognitive processes have often been inspired by new statistical tools. Thus, the scientific community would first have to rethink its routine use of multiple regression and similar techniques to facilitate accepting the idea that rational minds might not always weight but may simply tally cues. If this argument is correct, we will have to first teach better statistics to our future researchers in order to arrive at better models of mind.

1.2. Ignoring Cues to Make Better Predictions: Take-the-best

Consider again the canonical definition of rational inference as weighting and adding of all information (as long it is free). Yet, as illustrated in Figure 1-1, weighting and adding can lead to overfitting—that is, to excel in hindsight (fitting) but fail in foresight (prediction). The task of humans and other animals is to predict their world despite its inherent uncertainty, and in order to do this, they have to simplify. While tallying simplifies by ignoring the information required to compute weights, another way to simplify is to ignore variables (cues). The class of one-good-reason heuristics orders cues, finds the first one that allows a decision to be made, and then stops and ignores all other cues. Cues are ordered without paying attention to the dependencies between cues, but instead using a measure of the correlation between each cue and the criterion. This class of heuristic includes take-the-best (Gigerenzer & Goldstein, 1996), fast-and-frugal trees (Martignon, Katsikopoulos, & Woike,

2008), and the priority heuristic (Brandstätter, Gigerenzer, & Hertwig, 2006). For brevity, we focus here on take-the-best, but many of the results apply equally to the other heuristics.

The take-the-best heuristic is a model of how people infer which of two objects has a higher value on a criterion, based on binary cue values retrieved from memory. For convenience, the cue value that signals a higher criterion value is 1, and the other cue value is 0. Take-the-best consists of three building blocks:

1. Search rule: Search through cues in order of their validity.
2. Stopping rule: Stop on finding the first cue that discriminates between the objects (i.e., cue values are 1 and 0).
3. Decision rule: Infer that the object with the positive cue value (1) has the higher criterion value.

Take-the-best is a member of the one-good-reason family of heuristics because of its stopping rule: Search is stopped after finding the first cue that enables an inference to be made. Take-the-best simplifies decision making by both stopping after the first cue and by ordering cues unconditionally by validity, which for *i*th cue is given by:

$$v_i = \frac{\text{number of correct inferences using cue } i}{\text{number of possible inferences using cue } i}$$

Both these simplifications have been observed in the behavior of humans and other animals but routinely interpreted as signs of irrationality rather than adaptive behavior. In the late 1990s, our research group tested how accurately this simple heuristic predicts which of two cities has the larger population, using real-world cities and binary cues, such as whether the city has a soccer team in the major league (Gigerenzer & Goldstein, 1996, 1999). At the time, we were still influenced by the accuracy-effort trade-off view and expected that take-the-best might achieve a slightly lower predictive accuracy than multiple regression, trading off some accuracy for its simplicity. The unexpected result was that inferences relying on one good reason were more accurate than both multiple regression and tallying. We obtained the same result, on average, for 20 studies (Figure 1-1). This result came as a surprise to both us and the rest of the scientific community. In talks given at that time, we asked the audience how closely they thought the predictive accuracy of take-the-best would approximate multiple regression. The estimates of experienced decision theorists were consistently that take-the-best would predict 5 to 10 percentage points worse than multiple regression, with not a single guess in favor of take-the-best. For instance, the late decision theorist Ward Edwards was so surprised that he wrote an amusing limerick in response (published in Gigerenzer et al., 1999, p. 188).

But there were more surprises to come. Chater, Oaksford, Nakisa, and Redington (2003) used the city population problem and tested take-the-best against heavy-weight nonlinear strategies: a three-layer feedforward connectionist network, trained using the backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986); two exemplar-based models (the nearest-neighbor classifier [Cover & Hart, 1967] and Nosofsky's [1990] generalized context model); and the decision tree induction algorithm C4.5 (Quinlan, 1993). The predictive accuracy of the four complex strategies was rather similar, but the performance of take-the-best differed considerably. When the percentage of training examples (the sample size) was small or moderate (up to 40% of all objects), take-the-best outperformed or matched all the competitors, but when the sample size was larger, more information and computation seemed to be better. This was the first time that relying on one good reason was shown to be as accurate as nonlinear methods, such as a neural network. Yet, as Brighton (2006) showed in a reanalysis, Chater et al.'s method of fitting the models on the learning sample and then testing these models on the entire sample (including the learning sample) favored those models that overfit the data, especially at high sample sizes. When cross-validation was used, there was a new surprise: The predictive accuracy of take-the-best exceeded that of all rival models over the entire range of sample sizes (Figure 1-2). Cross-validation provides a far more reliable model selection criterion, and it is

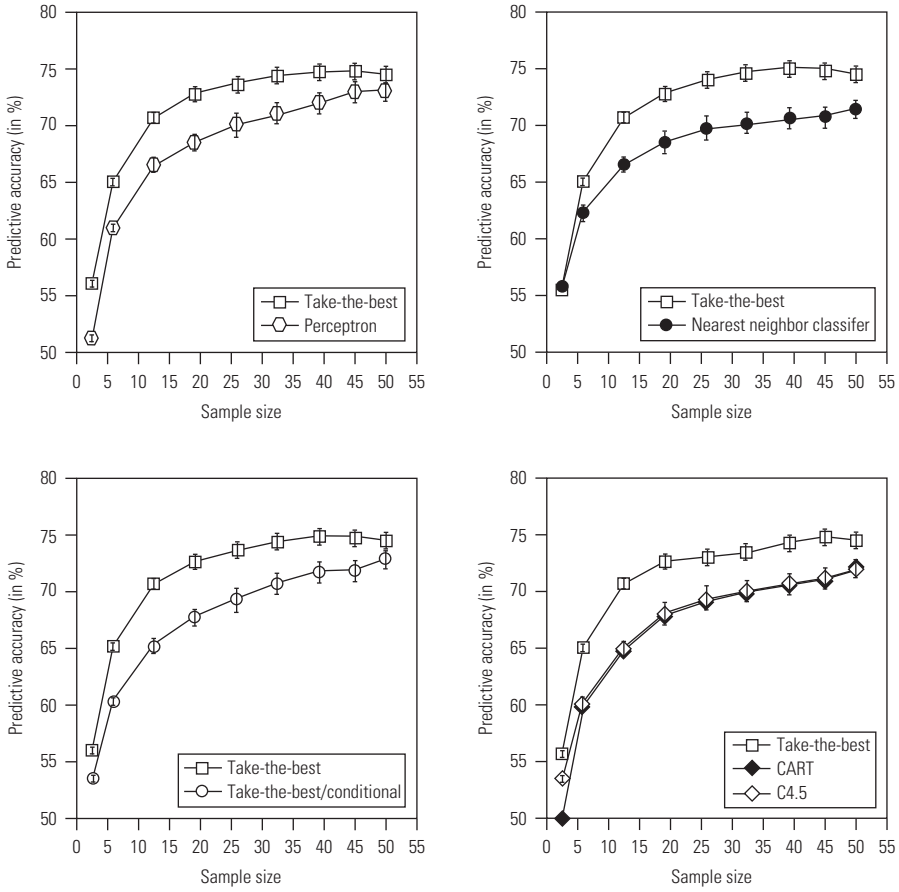

Figure 1-2. For the city population task, the performance of take-the-best was compared to five alternative models. Each panel plots the predictive accuracy of take-the-best and a rival model as a function of the number of objects used to train the model. Take-the-best outperforms (top left) a linear perceptron (essentially logistic regression); (top right) the nearest neighbor classifier; (bottom right) two tree induction algorithms, C4.5 and CART (classification and regression trees); and (bottom left) a variant of take-the-best that uses a more resource-intensive search rule that orders cues by conditional validity. Error bars are standard errors of means.

standard practice for assessing the relative performance of models of inductive inference (e.g., Hastie, Tibshirani, & Friedman, 2001; Stone, 1974). The same result was obtained for classification and regression trees (CART; Breiman, Friedman, Olshen, & Stone, 1984), which we thought might outperform take-the-best as a result of often inducing smaller decision trees than C4.5. An analysis of 20 data sets showed that the result in Figure 1-2 is the rule rather than the exception (Brighton, 2006).

Once again, another less-if-more effect was discovered, and a new question emerged: In which environments does relying on one good reason result in better performance than when relying on a neural network or on other linear and nonlinear inference strategies? We discuss this issue, the problem of understanding the ecological rationality of heuristics, in more detail in the next section. In short, the success of take-the-best seems to be due to the fact that it ignores dependencies between cues in what turns out to

be an adaptive processing policy when observations are sparse. Whereas all the competitors in Figure 1-2 attempt to estimate the dependencies between cues in order to make better inferences, take-the-best ignores them by ordering the cues by validity. In fact, when one alters the search rule of take-the-best by carrying out the more resource-intensive process of ordering cues by conditional validity, performance drops to the level of the more resource-intensive algorithms (Figure 1-2, bottom left). Conditional validity takes into account the fact that when one cue appears before another in the cue order, this first cue is likely to affect the validity of the second cue and all subsequent ones. Ordering cues by conditional validity is a costly operation, which requires recomputing the predictive value of cues against multiple reference classes of observations. This policy is similar to the process of recursive partitioning used by the machine-learning algorithms C4.5 and CART.

1.3. More Information, Computation, and Time Is Not Always Better

These two results are instances of a broader class of less-is-more effects found in the last decades, both analytically and experimentally. We use *less-is-more* here as a generic term for the class of phenomena in which the accuracy-effort trade-off does not hold, although the individual phenomena differ in their nature and explanation. The conditions under which the recognition heuristic (discussed below) leads to a less-is-more effect have been derived analytically (Goldstein & Gigerenzer, 2002) as have the conditions in which a beneficial degree of forgetting improves accuracy in inference (Schooler & Hertwig, 2005). Studies on language acquisition indicate that there are sensitive phases in which a reduced memory and simpler input ("baby talk") speeds up language acquisition (Elman, 1993; Newport, 1990); experiments with experienced handball players indicate that they make better decisions with less time (Johnson & Raab, 2003); and expert golfers (but not novices) do better when they have only 3 s to putt than when they can take all the time they want (Beilock, Bertenthal, McCoy, & Carr, 2004; Beilock, Carr, MacMahon, & Starkes, 2002). Like the surprising performance of equal weights, some less-is-more effects have long been known. For instance, in the (finitely repeated) prisoners' dilemma, two "irrational" players who play tit-for-tat can make more money than do two rational players who reason by backward induction and both defect (for an overview, see Hertwig & Todd, 2003; and Gigerenzer, 2007, 2008). The existence of these effects shows that there is more to heuristics than the accuracy-effort trade-off: The mind can use less information and computation or take less time and nevertheless achieve better performance.

Findings that show how less can be more have often been regarded as curiosities rather than as opportunities to rethink how the mind works. This is now changing. Most important, it has become clear that the discovery of less-is-more effects forces us to reassess our normative ideals of rationality. We turn now to the second step of progress made: the development of an understanding of *why* and *when* heuristics are more accurate than strategies that use more information and computation. The answer is not in the heuristic alone, but in the match between a heuristic and its environment. The rationality of heuristics is therefore ecological, not logical.

2. ECOLOGICAL RATIONALITY

All inductive processes, including heuristics, make bets. This is why a heuristic is not inherently good or bad, or accurate or inaccurate, as is sometimes believed. Its accuracy is always relative to the structure of the environment. The study of the ecological rationality asks the following question: In which environments will a given heuristic succeed, and in which will it fail? Understanding *when* a heuristic succeeds is often made easier by first asking *why* it succeeds. As we have shown, when analyzing the success of heuristics, we often find that they avoid overfitting the observations. For example, the ordering of cues chosen by take-the-best may not provide the best fit to the observations, but when predicting new observations, it often outperforms strategies that achieved a better fit. Indeed, all the models in Figure 1-2 achieved a better fit to the data than take-the-best did. The statistical

concept of overfitting is part of the explanation for why heuristics succeed, but to gain a clearer understanding of how and when heuristics exploit the structure of the environment, this issue can be examined more closely.

2.1. Heuristics and Bias

The study of heuristics is often associated with the term *bias*. The heuristics and biases program of Kahneman and Tversky used the term with a negative connotation: Reasoning errors reveal human biases that, if overcome, would result in better decisions. In this view, a bias is defined as the difference between human judgment and a "rational" norm, often taken as a law of logic or probability, such as statistical independence. In contrast to this negative use of bias, simple heuristics are perhaps best understood from the perspective of pattern recognition and machine learning, where there are many examples of how a biased induction algorithm can predict more accurately than an unbiased one (Hastie et al., 2001). Findings such as these can be explained by decomposing prediction error into the sum of three components, only one of which is bias:

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

The derivation of this expression can be found in many machine-learning and statistical inference textbooks (e.g., Alpaydin, 2004; Bishop, 1995, 2006; Hastie et al., 2001), but it is perhaps most thoroughly set out and discussed in a landmark article by Geman, Bienenstock, and Doursat (1992). The concepts of bias and variance can be understood by first imagining an underlying (true) function that some induction algorithm is attempting to learn. The algorithm attempts to learn the function from only a (potentially noisy) data sample, generated by this function. Averaged across all possible data samples of a given size, the bias of the algorithm is defined as the difference between the underlying function and the mean function induced by the algorithm from these data samples. Thus, zero bias is achieved if this mean function is precisely the underlying function. Variance captures how sensitive the induction algorithm is to the contents of these individual samples, and it is defined as the sum squared difference between the mean function, mentioned above, and the individual functions induced from each of the samples.

Notice that an unbiased algorithm may suffer from high variance because the mean function may be precisely the underlying function, but the individual functions may suffer from excess variance and hence high error. An algorithm's susceptibility to bias and variance will always depend on the underlying function and on how many observations of this function are available. The following example illustrates why seeking low bias will not always be functional for an organism, and how variance poses a significant problem when learning from finite data. The variance component of prediction error will prove essential to understanding why a less-is-more effect in heuristic inference can often be observed.

2.2. The Bias–Variance Dilemma

Figure 1-3A plots the mean daily temperature for London in 2000 as a function of days since January 1. In addition, the plot depicts two polynomial models attempting to capture the temperature pattern underlying the data. The first is a degree-3 polynomial model and the second a degree-12 polynomial model. As a function of the degree of the polynomial model, Figure 1-3B plots the mean error in fitting random samples of 30 observations. All models were fitted using the least squares method. Using goodness of fit to the observed sample as a performance measure, Figure 1-3B reveals a simple relationship: The higher the degree of the polynomial, the lower the error in fitting the observed data sample. The same figure also shows, for the same models, the error in predicting novel observations. Here, the relationship between the degree of polynomial used and the resulting predictive accuracy is U-shaped. Both high- and low-degree polynomials suffer from greater error than the best predicting polynomial model, which has degree-4.

This point illustrates the fact that achieving a good fit to observations does not necessarily mean we have found a good model, and choosing the model with the best fit is likely to result in

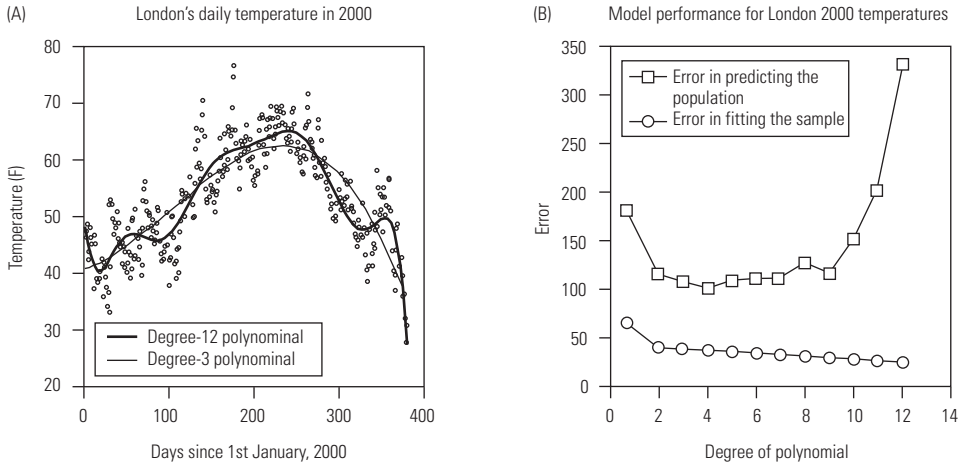

Figure 1-3. Plot (A) shows London's mean daily temperature in 2000, along with two polynomial models fitted with using the least squares method. The first is a degree-3 polynomial, and the second is a degree-12 polynomial. Plot (B) shows the mean error in fitting samples of 30 observations and the mean prediction error of the same models, both as a function of degree of polynomial.

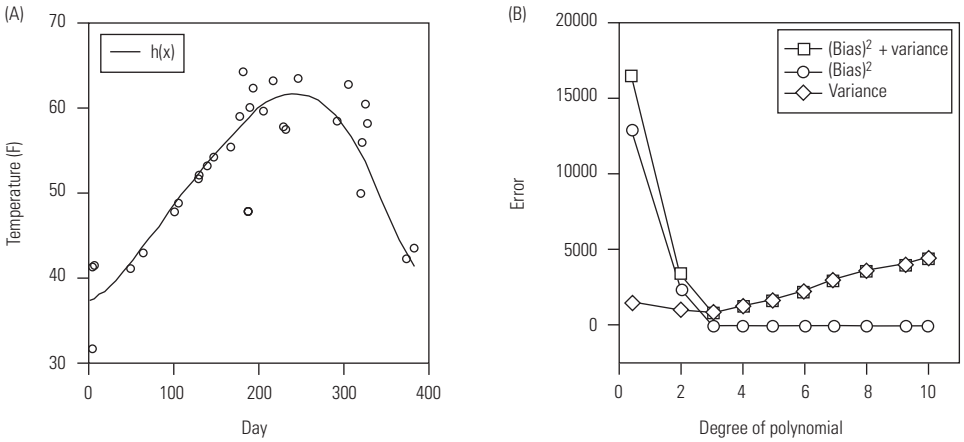

Figure 1-4. Plot (A) shows the underlying temperature pattern for some fictional location, along with a random sample of 30 observations with added noise. Plot (B) shows, as a function of degree of polynomial, the mean error in predicting the population after fitting polynomials to samples of 30 noisy observations. This error is decomposed into bias and variance, also plotted as function of degree of polynomial.

poor predictions. Despite this, Roberts and Pashler (2000) estimated that, in psychology alone, the number of articles relying on a good fit as the only indication of a good model runs into the thousands. We now consider the same scenario but in a more formal setting. The curve in Figure 1-4A shows the mean daily temperature in some fictional location for which we have predetermined the "true" underlying temperature pattern, which is a degree-3 polynomial, $h(x)$. A sample of 30 noisy observations of $h(x)$ is also shown. This new setting allows us to illustrate why bias is only one source of error impacting on the accuracy of model predictions.

The second source of error is variance, which occurs when making inferences from finite samples of noisy data.

As well as plotting prediction error as a function of the degree of the polynomial model, Figure 1-4B decomposes this error into bias and variance. As one would expect, a degree-3 polynomial achieves the lowest mean prediction error on this problem. Polynomials of degree-1 and degree-2 lead to significant estimation bias because they lack the ability to capture the underlying cubic function *h(x)* and will therefore always differ from the underlying function. Unbiased models are those of degree-3 or higher, but notice that the higher the degree of the polynomial, the greater the prediction error. The reason why this behavior is observed is that higher degree polynomials suffer from increased variance due to their greater flexibility. The more flexible the model, the more likely it is to capture not only the underlying pattern but unsystematic patterns such as noise. Recall that variance reflects the sensitivity of the induction algorithm to the specific contents of samples, which means that for different samples of the environment, potentially very different models are being induced. Finally, notice how a degree-2 polynomial achieves a lower mean prediction error than a degree-10 polynomial. This is an example of how a biased model can lead to more accurate predictions than an unbiased model.

This example illustrates a fundamental problem in statistical inference known as the bias–variance dilemma (Geman et al., 1992). To achieve low prediction error on a broad class of problems, a model must accommodate a rich class of patterns in order to ensure low bias. For example, the model must accommodate both linear and nonlinear patterns if, in advance, we do not know which kind of pattern will best describe the observations. Diversity in the class of patterns that the model can accommodate is, however, likely to come at a price. The price is an increase in variance, as the model will have a greater flexibility, which will enable it to accommodate not only systematic patterns but also accidental patterns such as noise. When accidental patterns are used to make predictions, these predictions are likely to be inaccurate. This is why we are left with a dilemma: Combating high bias requires using a rich class of models, while combating high variance requires placing restrictions on this class of models. We cannot remain agnostic and do both unless we are willing to make a bet on what patterns will actually occur. This is why "general purpose" models tend to be poor predictors of the future when data are sparse (Geman et al., 1992).

Our cognitive systems are confronted with the bias–variance dilemma whenever they attempt to make inferences about the world. What can this tell us about the cognitive processes used to make these inferences? First, cognitive science is increasingly stressing the senses in which the cognitive system performs remarkably well when generalizing from few observations, so much so that human performance is often characterized as optimal (e.g., Griffiths & Tenenbaum, 2006; Oaksford & Chater, 1998). These findings place considerable constraints on the range of potential processing models capable of explaining human performance. From the perspective of the bias–variance dilemma, the ability of the cognitive system to make accurate predictions despite sparse exposure to the environment strongly indicates that the variance component of error is successfully being kept within acceptable limits. Although variance is likely to be the dominant source of error when observations are sparse, it is nevertheless controllable. This analysis has important implications for the possibility of general-purpose models. To control variance, one must abandon the ideal of general-purpose inductive inference and instead consider, to one degree or another, specialization (Geman et al., 1992). Put simply, the bias–variance dilemma shows formally why a mind can be better off with an adaptive toolbox of biased, specialized heuristics. A single, general-purpose tool with many adjustable parameters is likely to be unstable and to incur greater prediction error as a result of high variance.

2.3. Explaining Less-is-more

The bias–variance dilemma provides the statistical concepts needed to further examine some of the less-is-more effects we have observed. First, 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 reproducing any response pattern that take-the-best can. Consequently, if a heuristic like take-the-best is to outperform an alternative method, it must do so by incurring less variance. Secondly, 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. This is why the performance of heuristic in an environment is not reflected by a single number such as predictive accuracy, but by a learning curve revealing how bias and variance change as more observations become available (Perlich, Provost, & Simonoff, 2003). Because the learning curves of two strategies—such as the pairs of curves in Figure 1-2—can cross, the superiority of one process over another will depend on the size of the training sample. 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.

To illustrate the point, we will perform a bias–variance decomposition of the error of take-the-best and a greedy version of take-the-best, which differs only in its use of a search rule that considers the cues in conditional validity order, discussed earlier (Martignon & Hoffrage, 2002). The following comparison between take-the-best and its greedy counterpart is insightful for two reasons. First, as Figure 1-2 suggests, the performances of the neural, exemplar, and decision tree models tend to be very similar to each other in paired comparison tasks, and 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. Second, Schmitt and Martignon (2006) proved that the greedy version of take-the-best is more successful in data fitting than take-the-best; yet when tested empirically, take-the-best nevertheless made more correct predictions (Martignon & Hoffrage, 2002). Comparing take-the-best and its greedy counterpart will allow us to examine when consuming fewer processing resources is likely to be functional, while also fixing the class of models used, as the two strategies are both restricted to inducing cue orders, rather than some richer class of models.

Two artificially constructed environments will be used to compare the strategies, both of which are 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. 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. Table 1-1 summarizes this result, along with two other studies that also aim to identify the environmental conditions favoring take-the-best. The second environment used in our comparison, however, is an instance of the class of *Guttman environments*, inspired by the Guttman scale (Guttman, 1944), where all the cues are maximally correlated with the criterion but vary in their discrimination rates. Formal definitions and illustrations of both these environments are provided in Appendix 1.

Figure 1-5A–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 important 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 point to the cases in which 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 to proving that the strategy also achieves a higher

Table 1-1. Three Examples of Environment Structure That Favor Take-the-best
The first two results are derived analytically and focus on the problem of achieving a good fit to the observations. From the perspective of bias and variance, these two results point to the cases in which take-the-best is unbiased. The third study addresses the case when performance refers to predictive accuracy and is based on simulation results and experiments with human participants.

Environment Structure	Result
Noncompensatory cue weights (Martignon & Hoffrage, 1999, 2002)	If (a) both the validities of the cues in take-the-best and the cue weights of a linear model have the same order; and (b) for each cue in this order, the weight of this cue is not exceeded by the sum of the weights of all subsequent cues; then (c) the inferences of take-the-best and the linear model will coincide.
Odds condition (Katsikopoulos & Martignon, 2006)	If (a) all cues are conditionally independent, and (b) an odds condition holds, then (c) take-the-best is optimal. Odds condition: When cues are ordered by validity, if the ith cue has validity v_i, then the odds condition holds if $\log(o_i) > \sum_{k>i} \log(o_k)$, where $o_i = v_i/(1-v_i)$.
Cue redundancy (Dieckmann & Rieskamp, 2007)	If cues are highly correlated and therefore carry redundant information, then take-the-best rivals or exceeds the predictive accuracy of a heuristic using the confirmation rule, described in the main text.

predictive accuracy. Third, and perhaps most important, Figure 1-5 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 1-2 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. More generally, as the variance component of error points to the sensitivity of a learning algorithm to the particular contents of samples, the essence of our argument does not hinge on the use of cross-validation as a model selection criterion. Adopting an alternative model selection criterion will not change the fact that the degree to which an algorithm is overly sensitive to the contents of the samples it learns from will depend on the size of the sample available. For example, very similar findings hold when the *minimum description length* principle is used as a model selection criterion (Brighton, 2006).

2.4. Biased Minds for Making Better Predictions

The relationship between mind and environment is often viewed from the perspective of bias, following the "mirror view" of adaptive cognition (Brighton & Gigerenzer, in press, set out the argument in more detail). In this view, a good mental model or processing strategy is assumed to be one that mirrors the properties of the world as closely as possible, preferably with no systematic bias, just as a linear model is assumed to be appropriate if the world is also linear. A cognitive system with a systematic bias, in contrast, is seen as a source of error and the cause of cognitive illusions. If this were true, how can cognitive heuristics that rely only on one good reason and ignore the rest make

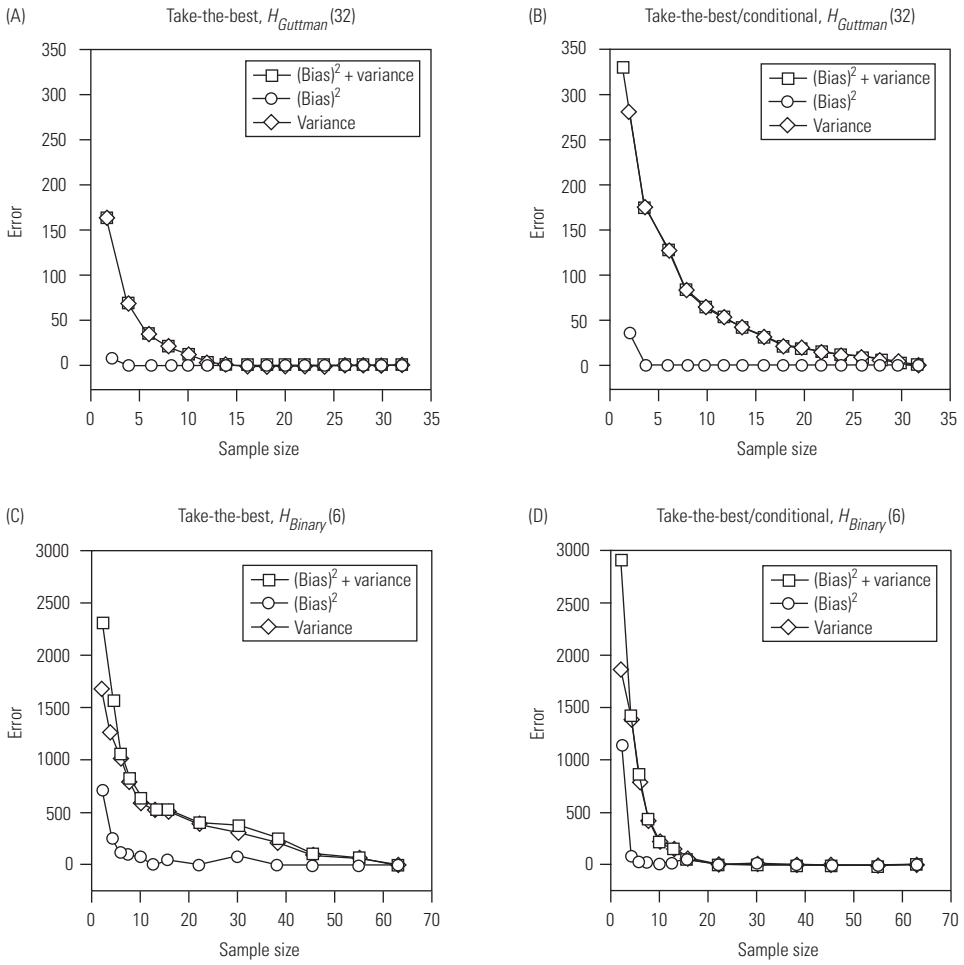

Figure 1-5. An illustration of the role played by variance in the performance of take-the-best. Plots (A) and (B) illustrate that, in Guttman environments, take-the-best outperforms the greedy variant of this heuristic, which orders cues by conditional validity. The performance difference is due to variance 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.

more accurate inferences than strategies that use more information and computation do (as illustrated in Figure 1-2)? We have identified three reasons:

1. The advantage of simplicity is not because the world is similarly simple, as suggested by the mirror view. This is illustrated by the apparent paradox that although natural environments exhibit dependencies between cues (such as the environment considered in Figure 1-2, where correlations between cues range between –.25 and .54), take-the-best can make accurate predictions by ignoring them so much so that it can outperform strategies that explicitly set out to model these dependencies. Superior performance is achieved by betting on lower variance, not lower bias.
2. As a consequence, if observations are sparse, simple heuristics like take-the-best are likely

to outperform more general, flexible strategies. It is under these conditions that variance will be the most dominant component of error.

3. Similarly, the more noise in the observations, the more likely a simple heuristic like take-the-best will outperform more flexible strategies. The greater the degree of noise, the more dominant the variance component of error is likely to be.

This argument is supported by a diverse set of related findings. First, consider how a retail marketing executive might distinguish between active and nonactive customers. Experienced managers tend to rely on a simple hiatus heuristic: Customers who have not made a purchase for 9 months are considered inactive. Yet there are more sophisticated methods, such as the Pareto/Negative Binomial Distribution model, which considers more information and relies on more complex computations. But when tested, these methods turned out to be less accurate in predicting inactive customers than the hiatus rule (Wübben & von Wangenheim, 2008). Second, consider the problem of searching literature databases, where the task is to order a large number of articles so that the most relevant ones appear at the top of the list. In this task, a "one-reason" heuristic (inspired by take-the-best) using limited search outperformed both a "rational" Bayesian model that considered all of the available information and PsychINFO (Lee, Loughlin, & Lundberg, 2002). Third, consider the problem of investing money into *N* funds. Harry Markowitz received the Nobel prize in economics for finding the optimal solution, the mean-variance portfolio. When he made his own retirement investments, however, he did not use his optimizing strategy, but instead relied on a simple heuristic: 1 /*N*, that is, allocate your money equally to each of *N* alternatives (see Table 1-2). Was his intuition correct? Taking seven investment problems, a study compared the 1 /*N* rule with 14 optimizing models, including the mean-variance portfolio and Bayesian and non-Bayesian models (DeMiguel, Garlappi, & Uppal, 2009). The optimizing strategies had 10 years of stock data to estimate their parameters, and on that basis had to predict the next month's performance; after this, the 10-year window was moved 1 month ahead, and the next month had to be predicted, and so on until the data ran out. 1 /*N*, in contrast, does not need any past information. In spite (or because) of this, 1 /*N* ranked first (out of 15) on certainty equivalent returns, second on turnover, and fifth on the Sharpe ratio, respectively. Even with their complex estimations and computations, none of the optimization methods could consistently earn better returns than this simple heuristic.

Finally, similar results in pattern recognition and machine learning include the finding that the naïve Bayes classifier can often outperform more sophisticated methods, even though its assumption of independence is explicitly violated by the task environment (Domingos & Pazzani, 1997). A related example is ridge regression, which can often outperform unbiased methods by introducing bias as a result of shrinking or ignoring beta weights, but can more than offset this error by incurring a greater reduction in variance (Hastie et al., 2001, p. 59).

Why should experts and laypeople rely on heuristics? To summarize, the answer is not simply in the accuracy-effort dilemma but in the bias–variance dilemma, as higher accuracy can be achieved by more or less effort. For instance, in predicting daily temperature (Figure 1-4), the relevant trade-off is between reducing the variance component of error by expending "less effort," that is, using a simpler polynomial, and reducing the bias component of error by expending "more effort," that is, using a more complex polynomial. The bias–variance dilemma is one of several principles that characterize the ecological rationality of heuristics (Gigerenzer et al., 1999; Hogarth & Karelaia, 2005, 2006; Katsikopoulos & Martignon, 2006; Todd, Gigerenzer, & the ABC Research Group, in press). The more uncertain the criterion is, or the smaller the sample size available, the more a cognitive system needs to protect itself from one kind of error (variance) over the other (bias). A biased mind that operates with simple heuristics can thus be not only more efficient in the sense of less effort but also more accurate than a mind that bets only on avoiding bias. The specific

Table 1-2. 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	Ecologically Rational If	Surprising Findings (examples)
Recognition heuristic (Goldstein & Gigerenzer, 2002)	If one of two alternatives is recognized, infer that it has the higher value on the criterion.	Recognition validity >.5	Less-is-more effect if $\alpha > \beta$; systematic forgetting can be beneficial (Schooler & Hertwig, 2005).
Fluency heuristic (Jacoby & Dallas, 1981)	If both alternatives are recognized but one is recognized faster, infer that it has the higher value on the criterion.	Fluency validity >.5	Less-is-more effect; systematic forgetting can be beneficial (Schooler & Hertwig, 2005).
Take-the-best (Gigerenzer & Goldstein, 1996)	To infer which of two alternatives has the higher value: (a) search through cues in order of validity, (b) stop search as soon as a cue discriminates, and (c) choose the alternative this cue favors.	See Table 1-1 and main text	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 & Karelaia, 2005, 2006)	Often predict equally or more accurately than multiple regression (Czerlinski et al., 1999).
Satisficing (Simon, 1955; Todd & 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 & Todd, 2002).	Aspiration levels can lead to significantly better choices than chance, even if they are arbitrary (e.g., the secretary problem, see Gilbert & Mosteller, 1966; the envelope problem, see Bruss, 2000).
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 & Goldstein, 2003; Pichert & Katsikopoulos, 2008)	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 for defection or cooperation but not divorce	Can lead to a higher payoff than optimization (backward induction).
Imitate the majority (Boyd & 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 & 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.

Note. For formal definitions, see references.

source of bias we identified lies in ignoring the dependencies between cues. The specific environmental structures that this bias can exploit are noisy observations and small sample sizes.

We turn now to a necessary precondition for understanding when, why, and how a given heuristic works.

3. FROM LABELS TO COMPUTATIONAL MODELS OF HEURISTICS

The development of computational models of heuristics over the last decades is another mark of progress, and without them, the analytic and simulation results on less-is-more could never have been discovered in the first place. Models need to be distinguished from labels. For instance, take similarity. On the one hand, there are models of similarity, including symmetric Euclidean distance and other Minkowski metrics (e.g., Shepard, 1974), as well as asymmetric similarity, such as Tversky's (1977) feature-mapping model. All of these are testable and some have been part of cognitive theories, such as Shepard's (1987) universal law of generalization. On the other hand, there is the label "representativeness," which was proposed in the early 1970s and means that judgments are made by similarity—but how similarity was defined is left open. A label can be a starting point, but four decades and many experiments later, representativeness has still not been instantiated as a model. As it remains undefined, it can account even for *A* and non-*A* (Ayton & Fisher, 2004), that is, everything and nothing. Consider the gambler's fallacy: After a series of *n* reds on the roulette table, expectation of another red *decreases*. This fallacy was attributed to people's reliance on the representativeness heuristic because "the occurrence of black will result in a more representative sequence than the occurrence of an additional red" (Tversky & Kahneman, 1974, p. 1125). Next consider the hot-hand fallacy, which is the opposite belief: After a basketball player scores a series of *n* hits, the expectation of another hit *increases*. This belief was also attributed to representativeness, because "even short random sequences are thought to be highly representative of their generating process" (Gilovich, Vallone, & Tversky, 1985, p. 295). No model of similarity can explain a phenomenon and its contrary; otherwise it would not exclude any behavior. But a label can do this by changing its meaning: To account for the gambler's fallacy, the term alludes to a higher similarity between the series of $n + 1$ outcomes and the underlying chance process, whereas to account for the hot hand fallacy, it alludes to a similarity between a series of n and observation number $n + 1$ (Figure 1-6). Labels of heuristics cannot be proved or disproved, and hence are immune to improvement.

In a debate over vague labels, Gigerenzer (1996) argued that they should be replaced by models, whereas Kahneman and Tversky (1996, p. 585) insisted that this is not necessary because "representativeness (like similarity) can be assessed empirically; hence it need not be defined a priori." The use of labels is still widespread in areas such as social psychology, and it is therefore worth providing a second illustration of how researchers are misled by their seductive power. Consider the "availability heuristic," which has been proposed to explain distorted frequency and probability judgments. From the very beginning, this label encompassed several meanings, such as the number of instances that come to mind (e.g., Tversky & Kahneman, 1973); the ease with which the first instance comes to mind (e.g., Kahneman & Tversky, 1996); the recency, salience, vividness, and memorability, among others (e.g., Jolls, Sunstein, & Thaler, 1998; Sunstein, 2000). In a widely cited study

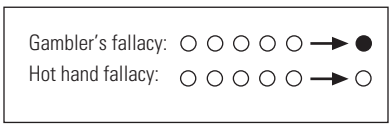

Figure 1-6. Illustration of the seductive power of one-word explanations. The gambler's fallacy and the hot hand fallacy are opposite intuitions: After a series of *n* similar events, the probability of the opposite event increases (gambler's fallacy), and after a series of similar events, the probability of the same event increases (hot hand fallacy). Both have been explained by the label "representativeness" (see main text).

designed to demonstrate how people's judgments are biased due to availability, Tversky and Kahneman (1973) had people estimate whether each of five consonants (K, L, N, R, V) appears more frequently in the first or the third position in English words. This is an atypical sample, because all five occur more often in the third position, whereas the majority of consonants occur more frequently in the first position. Two-thirds of participants judged the first position as being more likely for a majority of the five consonants. This result was interpreted as demonstration of a cognitive illusion and attributed to the availability heuristic: Words with a particular letter in the first position come to mind more easily. While this latter assertion may or may not be true, there was no independent measure of availability in this study, nor has there been a successful replication in the literature.

Sedlmeier, Hertwig, and Gigerenzer (1998) defined the two most common meanings of availability—ease and number—and measured them independently of people's frequency judgments. The number of instances was measured by the number of retrieved words within a constant time period (availability-by-number), and ease of retrieval was measured by the speed of the retrieval of the first word for each letter (availability-by-speed). The test involved a large sample of letters rather than the five consonants. The result was that neither measure of availability could predict people's actual frequency judgments. Instead, the judgments were roughly a monotonic function of the actual proportions, with a regression toward the mean, that is, an overestimation of low and an underestimation of high proportions. Moreover, the two definitions showed little correlation with one another. The basic flaw in the original research was to introduce a label after data were observed, rather than formulate a heuristic as a computational model that allows predictions to be expressed in a testable way.

One-word explanations such as the most recent "six general purpose heuristics identified (affect, availability, causality, fluency, similarity, and surprise)" (Gilovich & Griffin, 2002, p. 17) are not the only surrogates for theories. Others include redescription of the phenomenon[1] and so-called dual-process theories of thinking that postulate a System 1 and a System 2, each characterized by a list of general terms, such as rational versus heuristic, rule-based versus associative, and conscious versus unconscious, without formal and testable definitions of the two processes (e.g., Barbey & Sloman, 2007; Sloman, 1996; for a critique, see Gigerenzer & Regier, 1996). These surrogates for theories come close to black-box theorizing, and their popularity is an obstacle to progress in cognitive science.

In his essay "What is an 'explanation' of behavior?" Herbert Simon (1992, p. 155) wrote: "A running program is the moment of truth." A computational model can identify less-is-more effects, challenge the intuition that more information and computation are always better, and clarify which theories are psychologically implausible because they are computationally intractable in the real world. They also help us to derive, analytically, unexpected implications of a heuristic. For instance, violations of expected utility theory have been modeled by adding more and more adjustable parameters, as in cumulative prospect theory with its five adjustable parameters. The priority heuristic, a one-good-reason heuristic with no free parameters (Brandstätter, Gigerenzer, & Hertwig, 2008; Brandstätter et al., 2006) that has similar building blocks to take-the-best, has been shown to *imply* (not just have parameter sets that are *consistent with*) several of the major violations simultaneously, including the Allais paradox and the fourfold pattern (Katsikopoulos & Gigerenzer, 2008). The priority heuristic makes transparent and bold predictions about choice, and it has been demonstrated when it predicts well and when it does not (on the latter, see Birnbaum, 2008).

4. UNPACKING THE ADAPTIVE TOOLBOX

4.1. The Building Blocks of Heuristics

Although examples of rules of thumb in biology are not rare, they tend to be curiosities, partly because biology lacks a systematic theory of heuristics. Similarly, within cognitive science, there has been no such theory (although see Payne et al., 1993). The next step in progress we

deal with is the beginning of a systematic study of heuristics and their building blocks. Research into the adaptive toolbox attempts to formulate such a theory by identifying the heuristics that humans and other animals use, the building blocks of heuristics that can be used to generate new ones, and the evolved capacities that these building blocks exploit (Gigerenzer & Selten, 2001). The gaze heuristic introduced earlier has three building blocks. As pointed out, it only works when the ball is already high in the air, and fails if the ball is at the beginning of its trajectory. To adjust to this new situation, a player does not need a new heuristic, but only to adapt the third building block. Instead of

1. Fix your gaze on the ball,
2. start running, and
3. adjust your running speed so that the angle of gaze remains constant,

the adapted heuristic is:

1. Fix your gaze on the ball,
2. start running, and
3. adjust your running speed so that the image of the ball rises at a constant rate.

One can intuitively see its logic. If the player sees the ball rising from the point at which it was hit with accelerating speed, the player should run backward, because the ball will hit the ground behind the player's position. If, however, the ball rises with decreasing speed, the player needs to run toward the ball instead. Just as there is a class of such tracking heuristics, there is a class of one-good-reason heuristics, of which take-the-best is one member. These heuristics also have three building blocks: search rules, stopping rules, and decision rules. Take-the-best is not a useful heuristic in every situation; more generally, no single strategy is always the best one—otherwise, the mind would resemble a mechanic with only one tool at hand. Consider the first building block of take-the-best:

Search rule: **Search through cues in order of their validity.**

This search rule can be followed if the cues are retrieved from memory, but situations exist in which the order of cues is dictated from outside. Consider a red deer stag in rutting season that wants to enter the territory of a harem holder: In a fight with the rival over the females, which male is likely to win? For the stag, this question is a matter of genetic survival. Typically, the first cue is roaring. If the harem holder roars more impressively, the challenger may already give up and walk away. Otherwise, the next contest is initiated, parallel walking. It allows the competitors to assess each other's physical fitness and, potentially, confidence at a closer distance. If this contest also fails to produce a clear winner, the third contest is started: head-butting, the riskiest activity, as it can result in dangerous injuries (Clutton-Brock & Albon, 1979). This step-by-step heuristic is like take-the-best, but the search rule differs. The order of cues is not determined by (whatever the stag believes to be) the most valid cue, but by the cue that is first accessible. Sound can be encountered first in a forest environment where vision is restricted, visual stimuli next, and the most valid cue, head-butting, is last because it requires close contact. Thus, for the male deer, the adapted search rule is:

Search rule: **Search through cues in order of their environmental accessibility.**

The other building blocks remain the same. Consider now a situation where search by validity is not constrained. However, the task is new and the individual does not have the experience to come up with a good order. In this case, one can prove that it is of advantage to adjust the stopping rule and consequently, the decision rule (Karelaia, 2006):

Stopping rule: **Stop as soon as two cues are found that point to the same object.**
Decision rule: **Infer that this object has the higher criterion value.**

This stopping rule, termed a *confirmation rule*, works well in situations where (a) the decision maker knows little about the validity of the cues, and (b) the costs of cues are rather low (Karelaia, 2006). It is remarkably robust and insensitive to knowledge about cue ordering, and there is experimental evidence that a substantial proportion of people rely on this stopping rule as long as the problem is new (Gigerenzer, Dieckmann, &

Gaissmaier, in press). By adapting the building blocks of heuristics, organisms can react to new tasks and changing environments.

4.2. How Does the Mind Select Heuristics?

Table 1-2 shows 10 heuristics in the adaptive toolbox of humans. But how does the mind select a heuristic that is reasonable for the task at hand? Although far from a complete understanding of this mostly unconscious process, we know there are at least three selection principles. The first is that memory constrains the choice set of heuristics and thereby creates specific cognitive niches for different heuristics (Marewski & Schooler, 2010). Consider the choice between the first three heuristics in Table 1-2: the recognition heuristic, the fluency heuristic, and take-the-best. Assume it is 2003, and a visitor has been invited to the third round of the Wimbledon Gentlemen's tennis tournament and encouraged to place a bet on who will win. The two players are Andy Roddick and Tommy Robredo. First, assume that the visitor is fairly ignorant about tennis and has heard of Roddick but not of Robredo. This state of memory restricts the choice set to the recognition heuristic:

> If you have heard of one player but not the other, predict that the recognized player will win the game.

As it happened, Roddick won the match. In fact, this correct inference is not an exception: This simple heuristic predicted the matches of Wimbledon 2003 and 2005 with equal or higher accuracy than the ATP rankings and the seeding of the Wimbledon experts did (Scheibehenne & Bröder, 2007; Serwe & Frings, 2006). Now assume that the visitor has heard of both players but recalls nothing else about them. That state of memory limits the choice set to the fluency heuristic:

> If you have heard of both players, but the name of one came faster to your mind than the other, predict that this player will win the game.

Finally, assume that the visitor is more knowledgeable and can recall various facts about both players. That again eliminates the recognition heuristic and leaves a choice between the fluency heuristic and take-the-best. According to the experimental evidence, the majority of subjects switch to knowledge-based heuristics such as take-the-best when the values of both alternatives on relevant cues can be recalled (Marewski & Schooler, 2010), consistent with an analysis of the relative ecological rationality of the two heuristics in this situation. The general point is that memory "selects" heuristics in a way that makes it easier and faster to apply a heuristic when it is likely to yield accurate decisions. In the extreme case where the visitor has not heard of any of the players, none of the heuristics can be used. In this event, the visitor can resort to social heuristics, such as imitate the majority: Bet on the player on whom most others bet (Table 1-2).

The second known selection principle, after memory, is feedback. Strategy selection theory (Rieskamp & Otto, 2006) provides a quantitative model that can be understood as a reinforcement theory where the unit of reinforcement is not a behavior, but a heuristic. This model allows predictions about the probability that a person selects one strategy within a defined set of strategies. The third selection principle relies on the structure of the environment, as analyzed in the study of ecological rationality. For instance, the recognition heuristic is likely to lead to fast and accurate judgments if the recognition validity is high, that is, a strong correlation between recognition and the criterion exists, as is the case for tennis and other sports tournaments. There is experimental evidence that people tend to rely on this heuristic if the recognition validity is high but less so if the recognition validity α is low or at chance level ($\alpha = .5$). For instance, name recognition of Swiss cities is a valid predictor of their population ($\alpha = .86$), but not for their distance from the center of Switzerland, the city of Interlaken ($\alpha = .51$). Pohl (2006) reported that 89% of participants relied on the recognition heuristic in judgments of population, but only 54% in judgments of distance to Interlaken. Thus, the use of the recognition heuristic involves two processes: first, *recognition* in order to see whether the heuristic can be applied, and second, *evaluation* in order to judge whether it should be applied. Using functional magnetic resonance imaging (fMRI), Volz et al. (2006) reported specific neural activity that corresponded to these

two processes. Similarly, the take-the-best heuristic is more accurate when the weights of cues vary widely, but less so when they are about equal. Rieskamp and Otto (2006) and Bröder (2003) reported that people adaptively select take-the-best when the environment has this property.

5. METHODOLOGY AND EMPIRICAL EVIDENCE

Since the Nobel prize-winning experiments of ethologist Niko Tinbergen (1958), biologists have experimentally studied rules of thumb in settings such as mate choice, patch leaving, and the coordination of individual behavior of social insects, among others (Hutchinson & Gigerenzer, 2005). Yet this beautiful work has had virtually no influence on cognitive science. After the cognitive revolution of the 1960s, psychologists focused on heuristics for choice under certainty, such as how to decide between two or more apartments, contraceptive pills, or jobs, described by a number of cues or attributes. An early review of process-tracing studies concluded that there is clear evidence for noncompensatory heuristics, whereas evidence for weighting and adding strategies is restricted to tasks with small numbers of alternatives and attributes (Ford, Schmitt, Schechtman, Hults, & Doherty, 1989). A heuristic is noncompensatory if it makes no trade-offs between cue values; examples are lexicographic rules such as take-the-best, elimination-by-aspects, conjunctive rules, and disjunctive rules. The work by Payne et al. (1993) additionally demonstrated that people tend to select the heuristics in an adaptive way.

The experimental study of heuristics for inference began to attract researchers relatively late, after we published *Simple Heuristics That Make Us Smart* in 1999. An important stepping stone was realizing that a test of a heuristic should satisfy three minimum criteria:

1. *Competitive tests:* Test multiple models of strategies and determine which ones predict (rather than fit) the data most accurately. Do not test one model and declare that the result appears to be good enough or not.
2. *Individual-level tests:* Test each model for each individual. Do not test what the average individual does, because systematic individual differences may make the average meaningless.
3. *Adaptive selection of heuristics:* Test whether people use a heuristic in situations where it is ecologically rational. Do not test whether everyone uses one heuristic all the time.

Several tests of heuristics exist that satisfy these criteria. For instance, Bergert and Nosofsky (2007) formulated a stochastic version of take-the-best and tested it against an additive-weighting model at the individual level. They concluded that the "vast majority of subjects" (p. 107) adopted the take-the-best strategy. Nosofsky and Bergert (2007) compared take-the-best with both additive-weighting and exemplar models of categorization, and concluded that "most did not use an exemplar-based strategy" but followed the response time predictions of take-the-best. This work deserves particular admiration, given that Nosofsky long promoted exemplar models but has adapted his viewpoint in keeping with the evidence. Similar comparative tests of models of heuristics or their building blocks have been conducted by Bröder (in press), Bröder and Gaissmaier (2007), Dieckmann and Rieskamp (2007), Rieskamp and Hoffrage (2008), Rieskamp and Otto (2006), Yee, Dahan, Hauser, and Orlin (2007), among others.

Why is comparative testing crucial for progress? There are two reasons. First, science is about finding better models than those that already exist; the task is not to test one single model in isolation and then proclaim that it fits the data or does not. Consider Newell, Weston, and Shanks (2003), who tested take-the-best in two experiments and found that in Experiment 1 "participants adhered to the search, stopping, and decision rules in 75%, 80%, and 89% of all possible cases, respectively. For Experiment 2 the figures are even more impressive: 92% for the search rule and 89% for the stopping and decision rules" (p. 93). Yet these results were disregarded by counting only those subjects who were consistent with the predictions in at least 90% of the cases, so that "only one-third (33%)

behaved in a manner completely consistent with TTB's search, stopping, and decision rules" (p. 82). This and similar results were later used to argue that finding evidence for take-the-best "has proved elusive" (Newell, 2005, p. 12). Here, an arbitrary criterion is set unrealistically high for the model one wants to disprove, whereas a proper evaluation would be competitive by comparing several models.

A second reason why competitive tests are necessary is that flaws in the experimental design will hurt all models tested, not only one. For instance, in Newell and Shanks (2003), the ecological validities were set to .80, .75, .70, and .69, and subjects were given 60 trials with feedback to learn the order. The authors reported that only three out of 16 participants sought information in this order. But the training sessions were clearly too short to learn the order of cues: Even with perfect memory, a participant would have needed at least 100 trials to learn the order of the two last cues (to find out that the third cue is better than the fourth only in one out of 100 times), and 20 trials each for the other cues.

Individual-level tests are essential because in virtually every task we find individual differences in strategies. This heterogeneity may be due to flat maxima, where several strategies are reasonable solutions to the same problem, or a kind of Darwinian variability that is rational if the world (or task) changes, or a strategic unpredictability that can be rational in competitive games. As a consequence, models need to be tested at the individual level, whereas conclusions from group averages are likely to be uninformative. Consider early tests of the recognition heuristic, a model of memory-based inferences. If a person uses this heuristic to infer which of two alternatives has a higher value on a criterion, they rely only on recognition information and ignore cue values about the recognized alternative in their memory. Richter and Späth (2006) reported three experiments. The third was the perfect test for the heuristic, whereas in the other two experiments the recognition validity was unknown or low. German participants were taught whether certain recognized American cities have international airports. They were also told the proportion of cities in the reference class with airports, and that having an airport is highly predictive of population size. The task was to infer whether recognized cities with three kinds of associated information (airport, no airport, or no information) were more populous than unrecognized cities in paired comparisons.

The critical condition is the one where participants compare a recognized city with no airport and an unrecognized city. If participants follow the recognition heuristic, they will infer that the recognized city has the larger population, even if they know it has no airport. Because the mean adherence to the recognition heuristic was lower in the critical condition (.82) than in the other conditions (.95 and .98), Richter and Späth (2006, p. 159) concluded that "no evidence was found in favor of a noncompensatory use of recognition" but "clear evidence" for integration of recognition with additional knowledge. In Figure 1-7 we show a reanalysis of their data at the level of the individual participants. Even in the critical condition (Figure 1-7, bottom panel), the majority of participants consistently followed the recognition heuristic: nine participants followed it without a single exception (32 out of 32 times), and another eight followed it but for one single exception, whereas only a minority seemed to follow an alternative strategy. It is impossible to infer from means alone what proportion of individuals follow the recognition heuristic. Likewise, it is not possible to assert that there is clear evidence for an alternative noncompensatory model that has not even been formulated and tested. Marewski et al. (2010) formulated alternative models that integrate the additional information about the recognized information—the same models that Richter and Späth claimed to be supported by the fact that the recognition heuristic is not used all the time—and found that none of them were better than mere recognition at explaining the data. Similarly, individual-level analyses of several experiments showed that a majority of participants consistently followed the recognition heuristic in the presence of conflicting cues, while others employed, but less conistently, unidentified strategies (Pachur, Bröder, & Marewski, 2008).

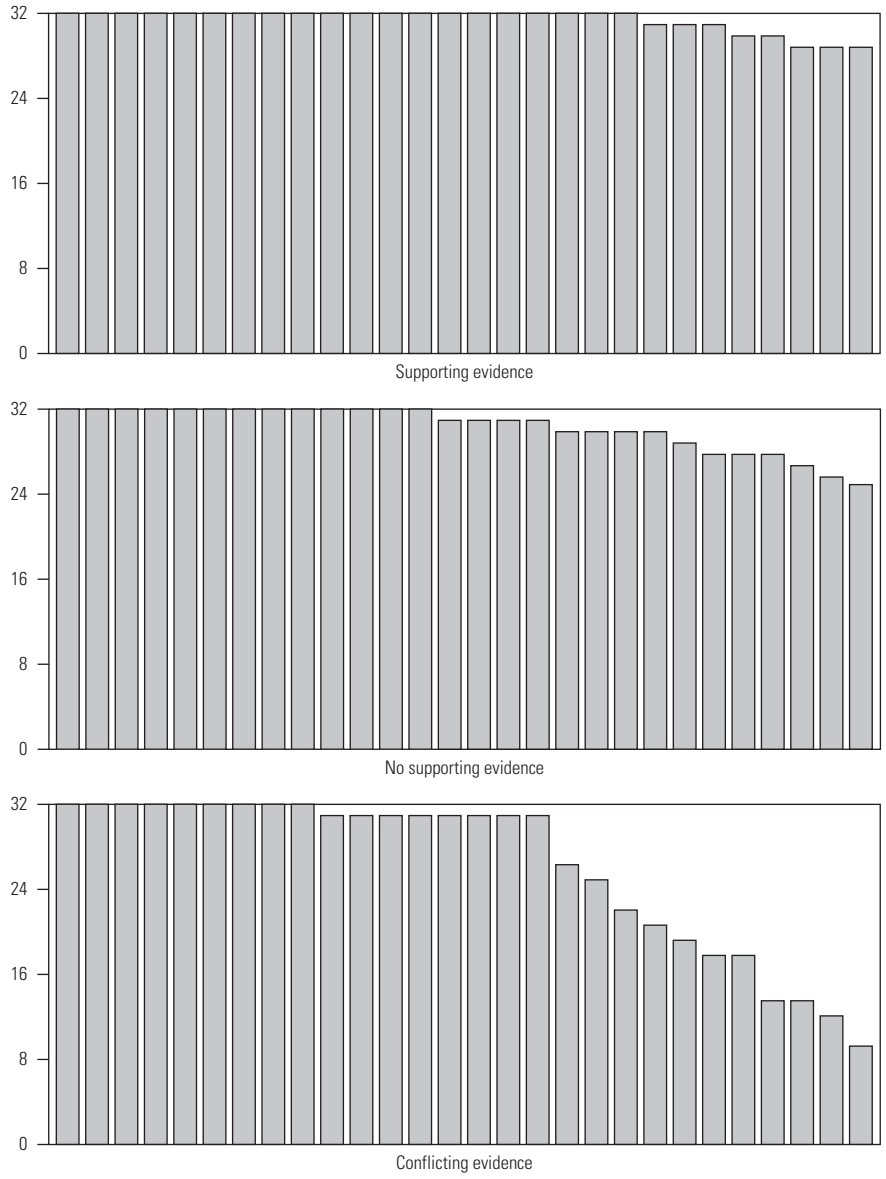

Figure 1-7. Systematic individual differences exist in the use of heuristics. A reanalysis of Richter and Späth (2006, Experiment 3) at the individual level reveals that even in the presence of conflicting information (bottom panel), the majority of participants still follow the recognition heuristic. Each bar represents one participant and its height the number of inferences out of a total of 32 that were consistent with the recognition heuristic. With supporting, no supporting, and conflicting evidence, the median participant followed the recognition heuristic in 100%, 97%, and 97% of inferences, respectively. We do not know which strategy the minority in the bottom panel (right side) followed, as no alternative strategies were tested. (Figure courtesy of Daniel G. Goldstein).

Finally, testing the adaptive selection of heuristics is essential to developing the theory of the adaptive toolbox as well as the earlier framework of the adaptive decision maker (Payne et al., 1993). As there is more than one heuristic, the question is not whether individuals always rely on a given heuristic, but whether they use heuristics in an adaptive way. Whereas some early papers set out to test whether people use a single heuristic all the time, later work asked whether people use several heuristics in an adaptive way (e.g., Bröder, in press). The study of ecological rationality of a heuristic provides the predictions for its adaptive use (e.g., Rieskamp & Otto, 2006).

6. HOMO HEURISTICUS

In this article, we presented a vision of human nature based on an adaptive toolbox of heuristics rather than on traits, attitudes, preferences, and similar internal explanations. We reviewed the progress made in developing a science of heuristics, beginning with the discovery of less-is-more effects that contradict the prevailing explanation in terms of accuracy-effort trade-offs. Instead, we argue that the answer to the question, "Why heuristics?" lies in their ecological rationality, that is, in the environmental structures to which a given heuristic is adapted. Using the bias–variance dilemma, we showed how the ecological rationality of heuristics can be formally studied, focusing on uncertain criteria and small samples that constitute environmental structures which fast-and-frugal heuristics can exploit. Homo heuristicus can rely on heuristics because they are accurate, not because they require less effort at the cost of some accuracy. We hope to have raised our readers' curiosity about the emerging science of heuristics and also hope that they might be inspired to solve some of the open questions, such as whether there is a system of building blocks of heuristics, similar to the elements in chemistry, and how a vocabulary for describing relevant environmental structures can be found. Let us end this article about the rationality of mortals from God's point of view.

How would a grand planner design a human mind? Consider three design perspectives. Design 1 would give the mind perfect memory. This would be ideal in a world that is absolutely certain and predictable, where what was observed in the past will also be observed in the future. This mind could remember, for instance, every day's temperature and thus could perfectly predict the future. In this world, perfect memory guarantees zero bias and zero variance, as every event has been observed and perfectly memorized. In fact, evolution has created something very close. A prominent example is a Russian named Shereshevsky, whom Luria (1968) studied for over three decades without finding the limits of his astounding memory. But this memory came at a price. For instance, Shereshevsky would read a page and recall it word for word, both forwards and backwards, but when he was asked to summarize the gist of what he read, he was more or less at a loss. Gist, abstraction, and other ways of going beyond the information given were not what this perfect memory buried in detail could deliver.

Design 2 accounts for the fact that the world is not absolutely predictable and fully observable, and therefore, a perfect memory would be a waste of energy. Instead, the goal is a mind that can make intelligent inferences from limited samples. The ideal is to have an infinitely flexible system of abstract representations to ensure zero bias, so that whatever structure the world has, it can be reproduced perfectly. As the content of the samples of observations in this world will vary, the induced representations are likely to be different, and this creates variance. Such a mind works best with large samples of observations and in a world that is relatively stable. Yet because this mind has no bias and must choose from an infinite space of representations, it is likely to require resource-intensive cognitive processing. This kind of design suggests general-purpose processing strategies such as exemplar models and neural networks as models of cognition (see Figure 1-2).

Design 3 aims at a mind that can make inferences quickly from a few observations, and it exploits the fact that bias can be adaptive and can help to reduce the estimation error (Figures 1-1 and 1-2). This design relies on several inference

tools rather than a single universal tool. Each has a bias that can be offset by a greater reduction in variance. This design works well in a world where inferences have to be made from small samples, and where the future may change in unforeseen ways (Bookstaber & Langsam, 1985). Unlike in the previous cases, the creator of Design 3 need not assume omniscience, that is, knowledge of all relevant options, consequences, and probabilities both now and in the future. This corresponds to the world in which many experts live, such as the business managers using the hiatus heuristic and the Nobel laureate relying on 1/*N*. After all, to build a mind with zero bias assumes that one knows the true state of the world and the representations needed to model it. Zero bias is neither possible nor always desirable for a real mind. Adopting the perspective of Design 3, the study of simple heuristics shows not only how a mind can make accurate inferences about an uncertain world efficiently but also how this efficiency is crucial to adapting the mind to its environment. Viewing humans as Homo heuristicus challenges widely held beliefs about the nature of cognitive processing and explains why less processing can result in better inferences.

ACKNOWLEDGMENTS

We are grateful to Julian Marewski, Shabnam Mousavi, Lael Schooler, and three anonymous reviewers for their helpful comments.

NOTES

1. Recall Moliere's parody of the Aristotelian doctrine of substantial forms: Why does opium make you sleepy? Because of its dormative properties. In the same way, explanation by redescription is alive today. Why does a specific problem representation improve reasoning? Because it makes the solution salient and transparent (for examples, see Gigerenzer, 1996, 2000). Why do people tend to share money equally in the ultimatum game? Because of a propensity for inequality avoidance (Fehr & Schmidt, 1999).

APPENDIX 1

Environments Used in the Bias–Variance Analysis

An environment is a collection of objects. Each object relates m binary cues to an integer criterion. The two classes of environment considered here are both parameterized by m. The first class of environment comprises *binary environments*, each of which has 2^m objects defined by

$$H_{\text{binary}}(m) = \left\{\left\langle b_m(i), i\right\rangle : 0 \le i \le (2^m - 1)\right\}$$

where the function b_m(i) maps integers onto their binary representations, coded using the binary cues (for example, $b_4(3) = (0, 0, 1, 1)$). Binary environments have noncompensatory weights, and the cues are uncorrelated. For example, $H_{\text{binary}}(3)$ defines the following environment:

Object	Cue1	Cue 2	Cue 3	Criterion
A	0	0	0	0
B	0	0	1	1
C	0	1	0	2
D	0	1	1	3
E	1	0	0	4
F	1	0	1	5
G	1	1	0	6
H	1	1	1	7

The second class of environment comprises *Guttman environments*, each of which has m objects given by

$$H_{\text{Guttman}}(m) = \left\{\left\langle b_m\left(\sum_{j=0}^{i} 2^j\right), i\right\rangle : 0 \le i \le (m-1)\right\}$$

For example, $H_{\text{Guttman}}(5)$ defines the following environment:

Object	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5	Criterion
A	0	0	0	0	1	0
B	0	0	0	1	1	1
C	0	0	1	1	1	2
D	0	1	1	1	1	3
E	1	1	1	1	1	4

In Guttman environments, all cues have validity 1.0 and are highly correlated with both other cues and the criterion. Binary and Guttman environments provide useful insights because they (a) elicit drastically different relative performances between take-the-best and alternative strategies that assess conditional dependencies between cues, and (b) are governed by a known underlying function, which allows us to perform a bias–variance decomposition of the error incurred by the different strategies.