9

Hindsight Bias

A Price Worth Paying for Fast and Frugal Memory

Ulrich Hoffrage
Ralph Hertwig

> Remembering is not the re-excitation of innumerable fixed, lifeless and fragmentary traces. It is an imaginative reconstruction, or construction . . .
>
> *Sir Frederic Bartlett*

Frustration about a fallible memory is familiar to most of us: "But men are men; the best sometimes forget" (Shakespeare, *Othello*). Remembering past events is not merely retrieving them from storage like books from a library. Memories can be lost or distorted, and memories for events that never even happened can be induced (e.g., Loftus, 1997; Schacter, 1995). Our memory is not like that of a Laplacean demon—we cannot perfectly recall everything we have ever thought, said, or experienced. Other chapters in this book deal with constraints of limited time and knowledge; in this chapter we focus on the constraints imposed by the limited capacity of human memory. How can memory work given its limitations? Our answer is, by *reconstruction*: When retrieval fails, inferential heuristics are employed. This answer is by no means new. It was already proposed by Sir Frederic Bartlett, one of the pioneers of modern memory research. In his classic *Remembering* (1932/1995, p. 213), Bartlett proposed that memory is a process of reconstruction (see—or recall if you can—the epigram that opened this chapter).

Reconstruction, however, has its price. We focus on one, the well-known hindsight bias, and propose a computational model for this effect based on a fast and frugal heuristic. Hindsight bias has often been regarded as just another error of human information processing. We argue, instead, that it is a by-product of two generally adaptive processes: first, updating knowledge after receiving new information; and second, draw-

ing fast and frugal inferences from this updated knowledge. Before specifying the model, we illustrate hindsight bias by exploring a topic that concerns every citizen of a modern democracy: public polling and electoral outcomes.

Public Polling, Elections, and Hindsight Bias

The history of American political polling is closely linked to the name of George Gallup. Gallup believed that his ideal of direct democracy called for public information and policy evaluation not filtered through the economic elite (Hamilton, 1995). In the early 1930s, he realized his vision of going directly to the voter by polling for a local election in Iowa in which his mother was a candidate. Shortly afterward, he began to apply this technique to predicting election results for dissemination by the public media. Since Gallup's early polls, polling has "moved to the epicenter of American campaigns" (Hamilton, 1995). For instance, Gallup and Harris, giants of the polling industry, attracted Richard Nixon's interest and became prime candidates for attack and manipulation by his administration (Jacobs & Shapiro, 1996). The acceptance of polling as a political tool did not go unchallenged. Indeed, it has been contended that opinion polls do not lead to political responsiveness, but are used by elites to manufacture the public attitudes they desire (see Jacobs & Shapiro, 1996).

The Achilles heel of polling companies is that the public can retrospectively check their predictions' accuracy. This is fine as long as they were accurate. In fact, for the Gallup company, the 1997 British parliamentary elections were just such a success story. In a poll sponsored by *The Daily Telegraph*, Gallup predicted the results almost perfectly. Based on interviews conducted a day before the elections with a randomly selected national sample of 1,810 citizens eligible to vote, Gallup forecast a 13% margin of victory for the Labour Party over the Conservative Party. The actual results of these historic elections, which ended the Conservatives' 18 years in power, put the final difference at 14%.

Such a post hoc reality check can, however, also be highly embarrassing. A famous "miss" by Gallup and others was Truman's victory in the 1948 presidential elections. In those first days of November, 1948, everyone knew that Thomas Dewey would defeat Harry Truman in the upcoming presidential elections. Pollsters and professional politicians alike predicted it. Daily newspapers came out eight to one in favor of Dewey. In its desire to get a scoop, the *Chicago Daily Tribune* jumped the gun in its November 4 edition and, relying on the seemingly reasonable predictions of Gallup and other polling companies, reported that Dewey would be the next president (Hamilton, 1995). (You may recall the photo of the smiling

president-elect, Harry Truman, holding aloft the newspaper with the now famous headline, DEWEY DEFEATS TRUMAN.)

What pollsters would like to do to save face in such situations is to say: "We knew it all along—that's what we really predicted." But the public memory represented in newspapers, videotapes, and other physical media make this ploy impossible—the pollsters must stand by their past predictions. Individuals speaking (or just thinking) in everyday life, however, usually have only their fallible internal memory to go on—and no external records to embarrass or contradict them. This can lead to situations in which an individual inaccurately remembers a prediction or statement he or she made in the past. For instance, Uncle Joe might contend that he knew Truman would win all along, even though he had earlier believed that Dewey would make it to the White House. This tendency to believe falsely—after the fact—that one would have predicted the outcome of an event is known as *hindsight bias* (for other systematic distortions in reconstructing the past, see Johnson & Sherman, 1990).

Recent laboratory research in psychology shows that hindsight bias is common in laypeople and experts (e.g., voters, physicians, businesspeople), and that it is manifest across a variety of judgments (e.g., confidence judgment, choice, categorization, or quantitative estimation; for a review, see Hawkins & Hastie, 1990). Not surprisingly, it also occurs in predictions of political election outcomes. For instance, before the 1982 Hawaiian gubernatorial election, Synodinos (1986) asked participants in a study to indicate the probability of each of the canditates winning the election. After the election, another group of participants was asked to make these predictions as if they had been asked before the election. As expected, the participants showed a "knew-it-all-along" tendency: The postelection probability estimates for the winner were higher than those made before the election, whereas the postelection estimates for the two losers were lower than the preelection estimates.

Synodinos (1986) demonstrated the effect of outcome knowledge by comparing two different groups of participants. Hindsight bias can also be found within a single participant. Fischhoff and Beyth (1975), for instance, had a group of student participants judge a variety of possible outcomes of President Nixon's visits to Peking and Moscow *before* they occurred in 1972. The possible outcomes were presented as assertions, such as: "The United States will establish a permanent diplomatic mission in Peking, but not grant diplomatic recognition" and "President Nixon will meet Mao at least once." Participants rated their confidence in the truth of the assertions on a 0% to 100% scale. *After* the visits, the assertions were repeated, and the participants were asked to recall their original confidence. The participants exhibited hindsight bias: Recalled confidence for events they thought had happened was higher than original confidence, while recalled confidence for events they thought had not happened was lower.

Views of Hindsight Bias

Hindsight bias has been interpreted in various ways. We distinguish two types of interpretations and add a third one. Fischhoff (1975), whose early experimental studies carved out this new topic for memory researchers, stressed that hindsight bias is not only robust and difficult to eliminate (Fischhoff, 1982a), but also has potentially harmful consequences:

> When we attempt to understand past events, we implicitly test the hypotheses or rules we use both to interpret and to anticipate the world around us. If, in hindsight, we systematically underestimate the surprises that the past held and holds for us, we are subjecting those hypotheses to inordinately weak tests and, presumably, finding little reason to change them. Thus, the very outcome knowledge which gives us the feeling that we understand what the past was all about may prevent us from learning anything from it. (Fischhoff, 1982b, p. 343)

Rather than stressing the harmful consequences of hindsight bias, others (e.g., Campbell & Tesser, 1983) have pointed out its potentially adaptive aspects. Presenting ourselves as wiser after the fact may enable us to appear intelligent, knowledgeable, or perspicacious. In fact, as long as no record of our previous judgments is available (which, unlike for pollsters, is generally the case), the immediate benefits of presenting oneself as knowledgeable outweigh the unlikely costs of being revealed as an imposter. In addition to hindsight's potential benefits in social interaction, hindsight bias may play an important role in creating and maintaining a coherent conception of oneself. Take, for instance, the situation of people who suddenly find themselves in a society whose value system has completely changed. The 1990s have seen an unusual number of such rapid societal transformations, from the fall of the apartheid regime in South Africa to the end of the socialist regimes in the Soviet Union and East Germany, among other countries. Many of those who held a responsible position in the old regimes are now being asked by their families and friends, or interrogated by official bodies (e.g., the Commission for Truth and Reconciliation in South Africa), to account for their previous behavior. Under these circumstances, hindsight bias—here the belief that one's past convictions and behavior are compatible with what the new regime considers to be right—can be an effective way of preserving the integrity of one's personality (and perhaps one's skin).

We propose a third view (which does not exclude the two views outlined above). According to this view, hindsight bias is a by-product of an adaptive process rather than being an adaptation itself (for a general version of this argument, see Campbell, 1959). To introduce this view, we first address the question: What are the alternatives to the assumption that human memory is unbounded in its capacity?

Consider the following situation. Mr. Loman is a salesman who visits

his clients by car. Every day, he repeatedly decides where to park his car, then stores this information in memory, and finally, after completing his business appointment, retrieves the car's location from memory. He does this very many times in the course of weeks, months, and years. How could a memory system be designed that allows Mr. Loman quickly and reliably to retrieve the knowledge about where he parked his car most recently? Is a system that maintains access to the knowledge of all past parking locations efficient? Some current conceptions of human memory seem to assume that we do in fact keep a record of every discrete event we have experienced and that, when we retrieve information or classify an object, we compare a probe with all our existing memory traces. For instance, exemplar models (e.g., Estes, 1986; Hintzman, 1988; Medin & Schaffer, 1978; Nosofsky, 1986; Ratcliff, 1978) are based on such an assumption. Although these models have provided impressive accounts of a wide array of memory phenomena, their psychological plausibility has been questioned, both for the extensive similarity computation, as well as for the vast memory resources they require (Nosofsky et al., 1994; see also chapter 11 for an alternative).

Sharing these doubts, we concur with Anderson and Schooler's (1991) argument that "it is just too expensive to maintain access to an unbounded number of items" (p. 396). In addition, a stockpile of memories (e.g., the memories of all the previous locations of Mr. Loman's car) may interfere with the only information that is relevant right now (e.g., where his car is currently parked). In this sense, forgetting may be necessary for memory to maintain its function, insofar as it prevents us from using old and possibly outdated information (Bjork, 1978; Ginzburg et al., 1996). A well-known phenomenon that reflects the adaptive nature of forgetting is the Zeigarnic effect. Zeigarnic (1927) showed that memory for tasks that have been completed declines rapidly compared to those tasks that have not yet been completed (e.g., a waiter's memory of the amount of the bill, depending on whether or not the customer has already paid). Thus, forgetting should most likely occur once the usefulness of some information has passed.

An alternative to a memory system that includes an immense, continuously expanding long-term storage is a system that maintains access primarily to the information most likely to be needed and most likely to be correct. For such a memory system, it is crucial to update information constantly and automatically. This process would avoid the problems of an exploding number of items, and the increasing retrieval time required if memory probes were compared with stored traces in a serial manner. It would make possible a boundedly rational memory system, which keeps available only those items that are most likely to be needed. Such a process of information updating is consistent with Bartlett's (1932/1995) classical finding that schemata are constantly changing and being updated.

Besides the fact that for most experiences there is no need for later recall (Anderson & Schooler, 1991), there is another reason why it is not

necessary to maintain a memory trace for everything we have thought, said, or experienced in the past: When something needs to be recalled, there are alternatives to memory retrieval. For example, imagine that you own 25 shares of stock in a company, which are listed in the newspaper as being worth $378.50 each. To calculate their total value, you multiply 25 times 378.5. A couple of days later, you want to know this value again. Can you remember $9,462.50? Probably not. However, this is not a problem, because you can compensate for your failure to retrieve it from memory by performing the same calculation again: Recall can be replaced by recalculation. We posit that the same sort of recalculation can be done—and, in fact, is done—when a past judgment, such as the prediction of the outcome of an election, needs to be recalled. If it cannot be recalled, going through the same process that led to the original judgment can provide a good approximation, and perhaps even a perfect substitute. There is, however, an important difference between a multiplication and a judgment. Performing arithmetic computations is a technical skill and we are trained to do it reliably. Therefore, performing the same multiplication a second time should yield the same result. In contrast, making a judgment often implies drawing knowledge-based inferences. If knowledge is constantly updated, as suggested above, inferences based on the updated knowledge may be different from those based on past knowledge.

Updating knowledge is the key assumption underlying the model of hindsight bias proposed below. It applies to situations where the original judgment was a knowledge-based inference. If the attempt to remember this original judgment directly fails, it will be reconstructed by repeating the same process that led to this judgment. However, knowledge about the outcome of an event, or feedback on whether an inference was correct, leads to an updating of relevant knowledge. As a consequence, the reconstruction based on the updated knowledge can be systematically different from the construction based on the original knowledge. This difference is what is known as hindsight bias. Thus, in our view, the so-called bias is a by-product of an adaptive process, namely knowledge updating.

Previously, we proposed a model that accounts for a puzzling effect in research on hindsight bias, namely the observation that hindsight bias is larger for assertions where the feedback is "true" than for assertions where the feedback is "false" (Hertwig et al., 1997). That model explained this finding as a result of the co-occurrence of hindsight bias and the reiteration effect, that is, the phenomenon that mere repetition of an assertion increases confidence in its correctness. However, that model does not explain why there is hindsight bias in the first place. The present model does. Although it is not the only account of hindsight bias where a hindsight judgment is seen as a "reconstruction of the prior judgment by 're-judging' the outcome" (Hawkins & Hastie, 1990, p. 321), it seems fair to say that ours is the only account that has specified a process model for knowledge-based inferences. It allows us to explain, at the level of

individual responses from individual participants, why hindsight bias occurred, did not occur, or even was reversed.

Inferring Past Judgments Fast and Frugally

What processes do people go through when they try to reconstruct their original judgment? We suggest that asking this question is the same as asking what processes underlie the original judgment. The theory of *probabilistic mental models* (PMM theory; Gigerenzer et al., 1991) provides one answer. The PMM framework applies to tasks in which a choice must be made between two alternatives according to a quantitative criterion, together with a judgment of confidence that the chosen alternative is correct. (In one such task, participants are asked: "Which city has more inhabitants, Heidelberg or Bonn?" "What is your confidence that the alternative you have chosen is the correct one?") We now extend the PMM framework to a context in which feedback about the correct answer is given, and the mind reconstructs the original response (both the choice and confidence). We call this model RAFT (for Reconstruction After Feedback with Take The Best).

Original Response

A concrete example will help to illustrate the task and the proposed mechanism: A friend of ours from southern California, Patricia, is trying to reduce her consumption of cholesterol. However, she has a sweet tooth and at a restaurant wants to order a dessert, either chocolate fudge cake or pumpkin custard pie. She asks herself which of the two foods has more cholesterol (in order to choose the one having less). Because Patricia does not know the correct answer, she tries to infer it from what she knows about the two foods. We hypothesize that to make this inference she will construct a probabilistic mental model. Such a PMM consists of a reference class, probability cues, knowledge about the objects of the reference class with respect to these cues, and a heuristic for processing this knowledge.

Knowledge About Cues According to PMM theory, knowledge is conceptualized as a set of cues (e.g., amount of saturated fat), and the values these cues have regarding the alternatives (henceforth, foods). When comparing the cue values of two foods, there are—in the case of a quantitative cue—four possible relations: "larger" (e.g., cake contains more saturated fat than pie), "smaller," "equal," or "unknown." Henceforth, we refer to these relations as *object relations.* (Note that it is sufficient to have an intuition, right or wrong, concerning the object relations; whether these

relations are directly retrieved or deduced from absolute cue values is left open.)

PMM theory also assumes that people have intuitions about the predictive power of a cue. The predictive power of a cue can be measured by its ecological validity. Ecological validity is defined as the relative frequency with which the cue correctly predicts which object scores higher on the criterion in a defined reference class (chapter 4). It is determined by considering only those comparisons where the cue discriminates (i.e., the object relation is "larger" or "smaller"). Let us assume that Patricia's reference class consists of foods sampled from her local supermarket, and let us consider saturated fat as a quantitative cue for cholesterol. When we took a random sample of 36 food items from a supermarket and checked all possible pairs, we found that in about 80% of these pairs, the food item with more saturated fat (cue) also has more cholesterol (criterion). This value is the ecological validity of the saturated fat cue (in our supermarket sample).

Heuristic How can Patricia use this knowledge to infer which food has more cholesterol? We account for her inference with a heuristic in the PMM framework called "Take The Best" (Gigerenzer & Goldstein, 1996a; see also chapter 4). If both foods are known, Take The Best starts with an estimated rank order of cues according to their validities and makes the inference on the basis of the highest ranking ("best") cue that discriminates between the two foods. Suppose that Patricia's PMM consists of three cues, amount of saturated fat, calories, and protein, which are already ordered according to their validities (80%, 70%, and 60%, respectively). Her original mental model about the relations between cake and pie on these cues is depicted in figure 9-1 (in the original response column). The highest ranking cue, saturated fat, does not discriminate; therefore Take The Best will try the next cue, calories. Because the cake has more calories than the pie, the heuristic stops searching and chooses cake as the alternative with more cholesterol. Confidence in the correctness of the decision is the validity of the cue that determined that decision (here, 70% as the validity of the calorie cue).

A defining characteristic of this fast and frugal heuristic is its simple stopping rule: Terminate search when the first good reason is found that speaks for one alternative over the other. No other cues are looked up after this point, and no cost-benefit computations are made in an attempt to determine the "optimal" stopping point for memory search. Such a simple stopping rule is crucial for memory-based inferences where much time and effort could be spent searching for information in the fog of memory. In the study reported below, we taught participants about only three cues (those listed in figure 9-1) and thus have artificially limited the search. In real-world inferences about food, there would typically be many more cues available and search would continue beyond such a small number of cues unless a stopping rule terminated it.

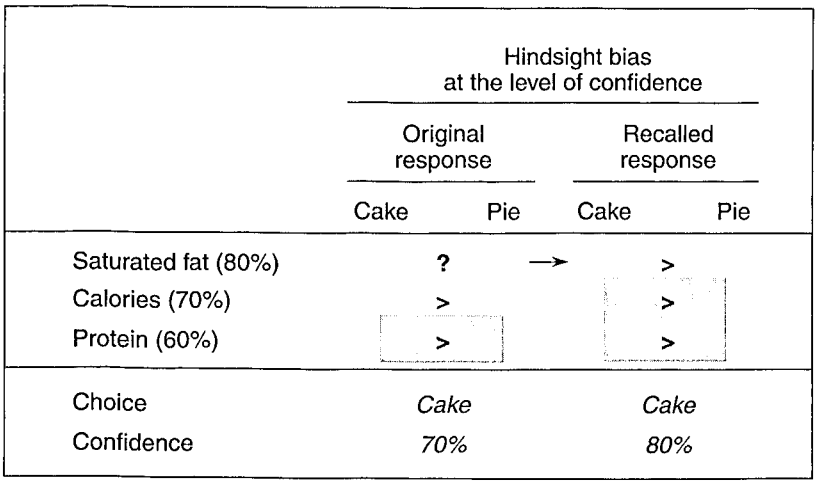


Figure 9-1: Hindsight bias at the level of confidence. The probabilistic mental model contains three cues ranked according to their validity (specified in parentheses). The symbols ">" and "?" denote the relations between objects on these cues. For instance, in the left column, which describes the knowledge underlying the original response, the object relation on the saturated fat cue is unknown. As indicated by the arrow ("→"), this object relation changes after feedback that cake has more cholesterol than pie. The relation shifts toward feedback, that is, from "?" to ">" in the updated mental model (right column). As a consequence, hindsight bias occurs. Note that Take The Best stops cue search before reaching the shaded object relations.

Feedback and Reconstruction

Some weeks after having dinner at the restaurant, Patricia goes to the market and finds out that chocolate fudge cake has more cholesterol than pumpkin custard pie. She tries to remember her past choice. What is the mechanism of recalling the original response? Figure 9-2 illustrates the cognitive processes as assumed by the RAFT model. First, an attempt is made to retrieve the original response directly from memory. The chance of doing this successfully depends on factors such as time delay between original judgment and recollection (Fischhoff & Beyth, 1975; Hertwig, 1996), and depth of encoding of the original response (Hell et al., 1988). If the original response is directly (and veridically) recalled from memory, no hindsight bias is obtained (upper left box in figure 9-2).

If the original response cannot be retrieved from memory, an attempt is made to reconstruct the original PMM that led to this response. An identical reconstruction will be obtained if (a) the type of strategy (e.g., lexicographic strategy, linear model, neural net, or Bayesian net) is the

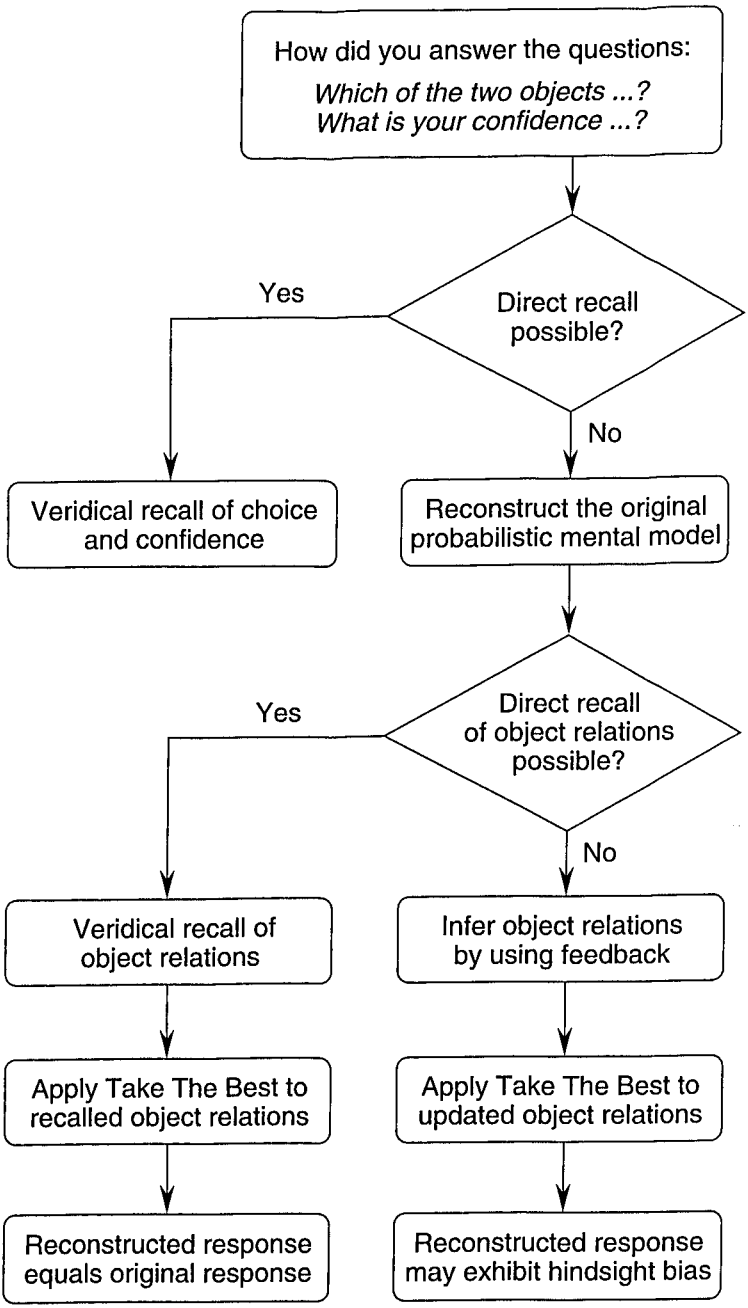


Figure 9-2: Flowchart of the RAFT model for hindsight bias.

same for the original response and its reconstruction; (b) this strategy operates with the same parameters (e.g., the same cue order, weights, or probabilities); (c) the strategy uses the same cues; and (d) the values that are retrieved on these cues are the same. A violation of any of these requirements may lead to differences between the original and the reconstructed response. In fact, RAFT posits a violation of requirement (d), that is, a systematic difference between the cue values underlying the original response and the reconstructed response. RAFT does not exclude the possibility that requirements (a), (b), and (c) may also be violated, and there are indeed such accounts of hindsight bias (e.g., Hawkins & Hastie, 1990). Nevertheless, we argue and provide evidence that the violation of requirement (d) is sufficient to account for hindsight bias.

Knowledge Updating Why should there be a change in object relations on the cues after feedback? Usually there is more than one cue that can be used to infer a criterion. Thus, if information on one cue is not available, this cue can be replaced by another. Egon Brunswik (1952) called this "vicarious functioning." Further, it is not only cues that are interchangeable, but also a cue and the criterion: For many cases, the possibility of drawing inferences from a cue to a criterion can also be reversed. For instance, not only can the amount of saturated fat be used to infer the amount of cholesterol, but the reverse is also true. Suppose you know neither how much saturated fat nor how much cholesterol is in chocolate fudge cake. If you now learn that cake has a lot of saturated fat, you can use this as a cue to infer that it also has a lot of cholesterol. Similarly, if you are told that cake has a lot of cholesterol, you can use this as a cue to infer a high saturated fat value. Thus, new information about the criterion can be used to update related knowledge in semantic memory—similarly to the updating of outdated information in episodic memory (as in Mr. Loman's car-parking case, see also Bjork, 1978). Updating is adaptive: It increases the coherence of our knowledge and the accuracy of our inferences (since more recent information is typically more valid and more relevant).

Thus, our conjecture is that knowledge stored in memory is in a state of flux, constantly changing, in part because new information is acquired, and in part because knowledge related to this new information is updated. Next, we show how such changes in knowledge over time lead us to predict and account for hindsight bias (as depicted in the branch at the bottom right of figure 9-2).

Predictions

The fact that the same variable may serve as either a criterion or a cue offers an interesting perspective. In the restaurant, Patricia worried about cholesterol, and her original mental model contained saturated fat as a cue (to infer cholesterol). In an attempt to reconstruct her original mental

model, saturated fat—or more precisely, the knowledge she had about saturated fat when she was in the restaurant—becomes the criterion. In some cases veridical retrieval of this past cue and current criterion may be possible; in others it may not. As a substitute for such a gap in memory, Patricia could use the knowledge she has now. However, in the meantime she found out which of the two foods has more cholesterol, and this new information might have led to an updating of related knowledge, such as saturated fat. As a consequence of this cue-criterion switch, the knowledge in the updated mental model will show systematic shifts toward feedback (Prediction 1). Because RAFT assumes that updating only occurs with some probability, this prediction does not necessarily hold for each single item.

According to RAFT, systematic changes in knowledge about cues can explain systematic changes in recollection of choice and confidence. That is, the occurrence of hindsight bias is contingent on the reconstructed knowledge. After excluding cases where original and recalled responses are identical (and thus can be attributed to direct memory), RAFT should be able to account for individual recollections: Regardless of whether hindsight bias or reversed hindsight bias is observed, this observation should match RAFT's prediction, which is derived from the recalled object relations for this item (Prediction 2).

Illustrations of the RAFT Model

We now illustrate how RAFT accounts for recollections made with hindsight. When Patricia tried to infer at the restaurant which of the two foods has more cholesterol, she did not know the value on the saturated fat cue (i.e., the most valid cue, see figure 9-1). After she found out that cake has more cholesterol than pie, this value was updated. The consequence is that in hindsight, when Patricia tries to remember her *original* judgment, the saturated fat cue—which was not available to her at the restaurant—discriminates, and Patricia infers that she thought that cake is the one with more cholesterol. She also infers that her confidence in this choice was 80% (the cue validity of saturated fat). Thus, her reconstructed choice is identical to her original choice. Her reconstructed confidence, however, increased relative to her original confidence. This is an example of hindsight bias. More generally, hindsight bias at the level of confidence occurs if a choice is correctly recalled, but recalled confidence increases after feedback indicating the originally selected alternative was correct (or decreases after feedback that it was wrong).

Not only can recalled *confidence* differ from original confidence, but recalled *choice* can differ from original choice as well. Hindsight bias at the level of choice occurs when feedback indicates that the originally chosen alternative was wrong and the recalled choice is the correct alternative (e.g., original choice is pie, feedback is cake, and recalled choice is cake). RAFT can account for hindsight bias at the level of choice as well.

Figure 9-3 (panel A) provides an example: At the restaurant, only the protein cue discriminated, pointing to the pie. After feedback, however, the saturated fat cue discriminates, pointing to the cake. If the original choice is reconstructed from this updated knowlegde, RAFT predicts hindsight bias at the level of choice.

It is also conceivable for hindsight bias to be reversed. Reversed hindsight bias at the level of choice occurs when the original choice is correct according to feedback, but recalled choice is wrong. Reversed hindsight bias at the level of confidence occurs if recalled choice equals original choice, but recalled confidence decreases after feedback confirming—or increases after feedback does not confirm—the originally selected alternative. How does RAFT account for reversed hindsight bias? It does so by allowing for random shifts in the object relations. That is, beyond systematic shifts due to feedback, RAFT posits unsystematic shifts due to the imperfect reliability of one's knowledge. Such random shifts are assumed to be independent of feedback. For this reason they may either coincide with the direction of feedback, or be counter to it. In figure 9-3 (panel B), a random shift changed the object relation on the saturated fat cue counter to the direction of the feedback. This random shift leads to reversed hindsight bias at the level of confidence.

To summarize, our starting point was the observation that human memory is bounded in its capacity. An alternative to unbounded memory is a system that maintains access to the information that is most likely to be needed and most likely to be correct. For such a memory system, it is

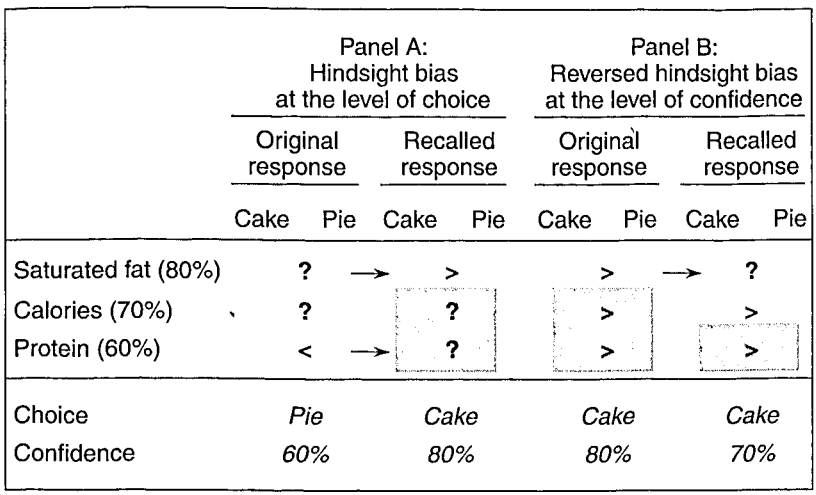


Figure 9-3: Hindsight bias at the level of choice (panel A), and reversed hindsight bias at the level of confidence (panel B). For an explanation of the symbols, see figure 9-1.

crucial to update information constantly and automatically. We suggest that hindsight bias is a by-product of this adaptive updating. Assuming a fast and frugal heuristic for reconstructing past judgments based on updated knowledge, RAFT explains this so-called bias.

Empirical Evidence

We conducted a study that was designed to test Predictions 1 and 2 (see Hoffrage et al., 1999). The experiment started with a phase in which participants learned the values of the saturated fat, calorie, and protein cues for various food items. They were also taught the validities of these cues (80%, 70%, and 60%) for inferring which of the food items has the higher amount of cholesterol. Immediately after this learning phase, the participants were given a list of food pairs and asked two questions about each pair: "Which food do you think has more cholesterol?" and "How confident are you that your choice is correct?" (The confidence rating scale ranged from 50% to 100%.) After they had given their responses, we asked them to recall the amounts of saturated fat, calories, and protein they had learned for each food item or to indicate for each food pair the relation between the food items on each cue (this is their knowledge before feedback).

In the second session, participants in the experiment first received the correct answer (feedback) for each of the questions they had answered in the first session. In the control condition, no feedback was provided. Then all participants were asked to recall (a) which food they had originally chosen as having more cholesterol, (b) how confident they were that their choice was correct, and—in a new questionnaire—(c) the originally learned cue values or the foods' relations on the cues (this is their knowledge after feedback). Recording participants' knowledge was important here because, according to RAFT, the occurrence of hindsight bias depends on this knowledge.

We first investigated whether participants showed any hindsight bias. For correct choices, hindsight bias occurred if recalled confidence increased, and for wrong choices hindsight bias occurred if recalled confidence decreased. In order to be able to include confidence judgments for correct and wrong choices in a single analysis, we mapped original and recalled confidences for wrong choices on a full-range scale. For example, a confidence judgment of 70% that the wrong alternative was the correct one was coded as 30% (confidence in the correct alternative). On this full-range scale, hindsight bias would always appear as an increase in confidence. Confidence increased in the feedback condition by an average of 3.4 percentage points, whereas in the no-feedback condition, it decreased by 0.6 percentage points. The effect of the difference is of medium size ($d = 0.54$, Cohen, 1988, p. 20) and is larger than the average effect size

reported in Christensen-Szalanski and Fobian Willham's (1991) meta-analysis.

Did relations on cues shift systematically after feedback (Prediction 1)? Shifts can occur toward or away from feedback. A shift toward feedback occurred when the cue originally pointed to the wrong alternative and now points to the correct alternative, or does not discriminate anymore. A shift toward feedback also occurred when the cue originally did not discriminate but now points to the correct alternative. The same logic defines shifts away from feedback. A cue does not discriminate if a participant did not specify the object relation on this cue (or the values for the two objects). In the feedback condition, 66.0% of the relations remained unchanged after feedback (across all participants, items, and cues). Did the remaining relations shift systematically toward feedback? Figure 9-4 shows the percentages of shifts toward and away from feedback. Consistent with Prediction 1, in the feedback condition, shifts toward feedback outnumbered those away from it, whereas in the no-feedback condition, both kinds of shifts occurred equally often.

Can we specify more precisely when shifts toward feedback occur? We suggest that updating after feedback should occur most likely when a cue did not discriminate at the time of the original response. To illustrate this rationale, let us first consider those cues that discriminated. The fact that a cue discriminated implies that knowledge was available in the original

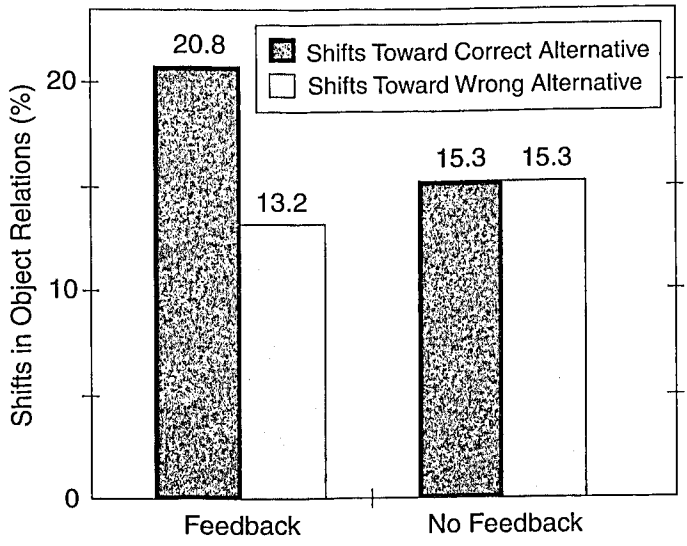


Figure 9-4: Proportion of object relations shifting toward and away from feedback. Shifts toward the correct or wrong alternative are equivalent to shifts toward or away from feedback, respectively, when feedback is given.

mental model. The mere existence of knowledge provides the chance that it can be accessed again at some later point and that the object relation is veridically retrieved—even after feedback. In contrast, if the relation was unknown, then feedback does not need to overcome preexisting knowledge to become manifest. A similar implication holds for "equal" relations. Here, a shift in one cue value is sufficient to change the relation. For a discriminating relation, a shift may reduce the difference between the two values but not necessarily cause a change in the relation.

Is updating after feedback most likely when a cue did not discriminate at the time of the original response? To answer this question, we calculated the differences in the proportions of shifts toward and away from feedback (across all participants, items, and cues). Figure 9-5 displays the results. In 38.7% of the cases in which cues originally did not discriminate, cues disciminate after feedback: in 27.7% of the cases, they point to the correct alternative, and in 11.0% to the wrong alternative. This difference of 16.7 percentage points is depicted by the leftmost bar in figure 9-5. In contrast, when cues originally discriminated, shifts were almost symmetrical—the difference between shifts toward and away from feedback decreased to 2.8 percentage points. In the no-feedback condition, the difference between shifts was miniscule for both discriminating and nondiscriminating cues. These results strongly confirm the prediction that the impact of feedback is most pronounced when a cue did not discriminate at the time of the original response.

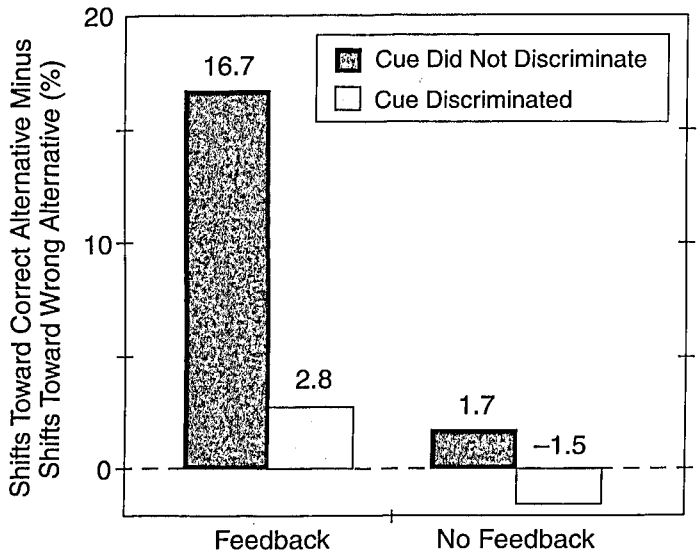


Figure 9-5: Proportion of object relations shifting toward the correct alternative minus those shifting toward the wrong alternative, depending on whether a cue discriminated when the original response was given.

Can hindsight bias and reversed hindsight bias be predicted from recalled object relations (Prediction 2)? To test this prediction, we first excluded cases where original and recalled response are identical, because they can be attributed to accurate memory (and are thus not subject to reconstruction). Next, we determined RAFT's prediction for each participant as follows. For each food pair, we applied Take The Best to the updated knowledge and compared the resulting choices and confidences with the original choices and confidences. This comparison determined whether RAFT would predict hindsight bias, reversed hindsight bias, or no hindsight bias (predicted outcome). The predicted outcome was then compared with observed outcomes (hindsight bias or reversed hindsight bias), and for each participant, we finally determined the percentage of correct predictions across items.

Averaged across all participants, the percentage of correct predictions was 76.3%. RAFT correctly explains nearly as many of the observed outcomes in the feedback and the no-feedback conditions: 76.6% and 75.9%, respectively. This is not surprising as RAFT can also account for reconstructed judgments based on cue values that were *not* updated (e.g., because *no* feedback was provided). To see how good the performance of RAFT is we compared it with a chance model (for details, see Hoffrage et al., 1999). Averaged across all participants in the feedback and the no-feedback conditions the performance of this chance model was 67.9% (i.e., 8.4 percentage points worse than RAFT's performance; $t = 5.0$, $p = .001$).

Looking Back

We proposed a model of the cognitive processes underlying hindsight bias. This model assumes that information about the correct answer leads to an updating of elusive cue values. If the original response is inaccessible, it will be reconstructed based on cue values that may have been updated. As a consequence, the reconstructed response may exhibit hindsight bias. Consistent with Prediction 1, we found that feedback on the criterion systematically influenced participants' recollection of their knowledge about cues. Consistent with Prediction 2, a majority of the cases in which either hindsight bias or reversed hindsight bias occurred was accurately predicted by applying Take The Best to the recalled (and updated) cue values. In Hoffrage et al. (1998), we report a further study that replicated the present results, report evidence for a third prediction (that assisting the recall of cue values reduces hindsight bias), and discuss how RAFT explains other findings obtained in research on hindsight bias.

Fast and Frugal Inferences

The model of hindsight bias we have proposed integrates ideas from Sir Frederic Bartlett, Egon Brunswik, and Herbert Simon. Like Bartlett, we

see remembering as a process of reconstruction. Bartlett himself did not go on to specify how this reconstruction can be modeled. In our view, consistent with Brunswik's (1952, 1957) framework, reconstruction is based on uncertain cues. However, in contrast to the neo-Brunswikian idea that cues are weighted and integrated by multiple regression (Cooksey, 1996; Doherty, 1996; Hammond, 1955), our assumption is that the nature of the inferential mechanism is fast and frugal. Take The Best is such a fast and frugal mechanism. Because it has a stopping rule, it does not seek all the available information, and it is computationally simple compared with multiple regression. Thus, RAFT's inferential mechanism is a bounded rational one (Simon, 1982; chapter 1).

As many of the results reported in this book suggest, Take The Best can compete impressively with more complex strategies. Because of its psychological plausibility, we chose to model people's recollections with this simple heuristic. Would we have achieved a better fit of the predicted and actual responses if we had used computationally more powerful but psychologically less plausible strategies? To answer this question, we reanalyzed the data and tested Prediction 2 with several other strategies, including a unit-weight linear model (Dawes's rule, chapter 4), a linear model with cue validities as the weights (Franklin's rule, chapter 4), and naive Bayes (chapter 8). None of the alternative strategies modeled actual responses better than Take The Best; they all performed similarly well. One reason for this is that the strategies' responses were generated from only three cues, and thus for most constellations of cue values they made the same inference (see chapters 6 and 7). As it is psychologically more plausible, and in light of the evidence that people's choices can best be modeled by heuristics that only process some of the available information (chapter 7), we suggest Take The Best as the more likely candidate strategy of people's memory inferences.

Conclusion

The adaptive process of knowledge updating relieves us of the need to store everything we have thought, said, or experienced in the past. Updating makes us smart by preventing us from using information that may be outdated due to changes in the environment. As Bartlett put it: "In a world of constantly changing environment, literal recall is extraordinarily unimportant" (1932/1995, p. 204). Adaptive updating has an uninvited by-product: hindsight bias. But this by-product may be a relatively low price to pay for a memory that works fast and frugally.

10

Quick Estimation

Letting the Environment Do the Work

Ralph Hertwig
Ulrich Hoffrage
Laura Martignon

> We may look into that window [on the mind] as through a glass darkly, but what we are beginning to discern there looks very much like a reflection of the world.
>
> *Roger N. Shepard*

"*September 30, 1659.* I, poor, miserable Robinson Crusoe, being shipwrecked, during a dreadful storm in the offing, came on shore on this dismal unfortunate island, which I called 'the Island of Despair,' all the rest of the ship's company being drowned, and myself almost dead" (Defoe, 1719/1980, p. 74). Thus begins *Robinson Crusoe.* Daniel Defoe's classic novel has been interpreted as everything from a saga about human conquest over nature to an allegory about capitalism. At a much more mundane level, however, Crusoe's adventures illustrate the crucial importance of being able to estimate the frequency of recurrent natural events accurately. Of his first attempt to sow grain, he wrote in his journal: "Not one grain of that I sowed this time came to anything; for the dry months following, the earth having had no rain after the seed was sown" (p. 106). From then on, Crusoe kept track of the rainy and dry days in each month, and subsequently sowed seed only when rainfall was highest. He reaped the rewards of this strategy, later reporting: "I was made master of my business, and knew exactly when the proper season was to sow; and that I might expect two seed times, and two harvests, every year" (p. 107).[1]

1. Crusoe's story may not be completely fictitious. Before the publication of *Robinson Crusoe,* Defoe might have read about Alexander Selkirk, a sailor who survived five years on a desert island—Juan Fernandez Island off the coast of

Defoe equipped the fictional Crusoe with a journal, which helped him to predict rainfall. Are real humans equipped to estimate environmental quantities even without the benefit of written records? One domain where we would expect to find evidence of such an ability—if it exists—is in foraging for food. Humans have spent most of their evolutionary history in hunter-gatherer foraging economies in which they have had to decide what to hunt. The Inujjuamiut, a group of Eskimos who live in Canada, afford us an opportunity to observe how contemporary human hunter-gatherers select strategies for obtaining food (Smith, 1991). One of the Inujjuamiuts' food sources is the beluga whale. When hunting belugas, the Inujjuamiut encircle a group of them and drive them into shallow water. Exploiting the whales' sensitivity to noise, the hunters then "herd" them by pounding on the gunwales of their canoes and shooting in a semicircle around them. While the whales are being killed with high-powered rifles and secured with floats, the pursuit of the next group of belugas gets underway.

Inujjuamiut foraging strategies—their strategies for choosing prey and hunting methods—can be modeled by the *contingency prey model.* According to the anthropologist Eric Alden Smith (1991, p. 237), this model is the best tool yet devised for explaining hunter-gatherer prey choice. It suggests why the Inujjuamiut undertake time-consuming and dangerous whale hunts rather than pursuing easier prey, such as ducks, geese, and seals. Its basic intuition, shared by other foraging models, is that a forager who has encountered a food item (prey or patch) will only attempt to capture it if the return per unit time for doing so is greater than the return that could be obtained by continuing to search for another item. Hence prey choice depends on rankings of food items in terms of return rates (see chapter 15). Setting aside the details of this model (see Smith, 1991), one of its crucial assumptions is that to be ranked according to their net return, food items (from prey) must be classified according to their statistically distinct *return rates* (per-unit handling time, i.e., time spent in pursuit, capture, and processing) and *encounter rates* (per-unit search time). Thus, just as Defoe equipped Crusoe with journal entries from which to estimate rainfall, the contingency prey model endows humans with the cognitive abilities necessary to estimate environmental quantities (e.g., the rate at which they encounter a certain type of prey).

But literary devices and theoretical assumptions aside, the question remains: Do humans actually have this ability, and how can it be modeled? According to Brown and Siegler (1993), psychological research on real-world quantitative estimation "has not culminated in any theory of esti-

---

Chile. Selkirk was left there at his own request after quarreling with his captain. When it was published, Selkirk's story was a sensation. The public was fascinated by the way this man had survived—as was Defoe, who may even have met him, as some scholars believe (see Swados's Afterword in Defoe, 1719/1980).

mation, not even in a coherent framework for thinking about the process. This gap is reflected in the strangely bifurcated nature of research in the area. Research on heuristics does not indicate when, if ever, estimation is also influenced by domain-specific knowledge; research on domain-specific knowledge does not indicate when, if ever, estimation is also influenced by heuristics" (p. 511). In this chapter, we attempt to bridge this gap by designing a heuristic adapted to make fast and frugal estimates in environments with a particular statistical structure. Before describing this heuristic, we review previous research on quantitative estimation, focusing on how people estimate numbers of events (both types and tokens); the events in question may be objects, people, or episodes.[2] We review two classes of estimation mechanisms: estimation by direct retrieval and estimation by inference.

Estimation by Direct Retrieval

The Scottish Enlightenment philosopher David Hume believed that the mind unconsciously and automatically tallies event frequencies and apportions degrees of belief in events accordingly. Hume (1739/1975) claimed that the psychological mechanism for converting observed frequency into belief was extremely finely tuned: "When the chances or experiments on one side amount to ten thousand, and on the other to ten thousand and one, the judgment gives the preference to the latter, upon account of that superiority" (p. 141).

Recent research on human monitoring of event frequencies (Hasher & Zacks, 1979, 1984) supports Hume's position by suggesting that memory is extremely sensitive to frequency of occurrence information (Hasher & Zacks, 1984, p. 1379), although not as finely tuned as Hume suggested. People's sensitivity to natural frequency of occurrence has been demonstrated using a variety of stimuli. For instance, several authors have documented that people's judgments of the frequency with which letters and words occur generally show a remarkable sensitivity to their actual frequencies (e.g., Attneave, 1953; Hock et al., 1986; Johnson et al., 1989).[3]

Hasher and Zacks (1979, 1984) assumed that people automatically encode the occurrences of an event, store a fine-grained count of its frequency, and when required to estimate its frequency, access this count. They proposed that people can estimate frequencies accurately because

2. This chapter does not review research on estimation of psychophysical stimuli (e.g., Haubensack, 1992; Mellers & Birnbaum, 1982; Parducci, 1965), probabilities (e.g., Kahneman et al., 1982; Peterson & Beach, 1967), or statistical parameters, such as central tendency, variability, and correlation (e.g., Busemeyer, 1990).

3. For instance, Attneave (1953) asked participants to judge the relative frequencies of all the letters in the alphabet and found a correlation of .79 between actual relative frequencies and the medians of the judged frequencies.

registering event occurrences is a fairly automatic process, that is, it requires little to no attentional capacity. In this view, frequency is one of the few attributes of stimuli that seems to be encoded automatically (others being spatial location, temporal information, and word meaning). Although the claim that event frequencies are automatically encoded may be too strong and has been seriously criticized (see Barsalou, 1992, chap. 4), there seems to be broad agreement with the conclusion that Jonides and Jones (1992) summarized as follows: "Ask about the relative numbers of many kinds of events, and you are likely to get answers that reflect the actual relative frequencies of the events with great fidelity" (p. 368). A similar conclusion has also been drawn in research on probability learning, about which Estes (1976) remarked: "The subjects clearly are extremely efficient at acquiring information concerning relative frequencies of events" (p. 51).

Estimation by Inference

Where Hasher and Zacks assume that people have access to a count of the event, the advocates of a rival approach contend that people infer this value from cues correlated with it. The researchers who advocate this approach may be divided into two groups according to their postulate of the nature of these cues: ecological versus subjective.

Inference by Ecological Cues

According to Brunswik (1952, 1955), the perceptual system estimates a distal variable (e.g., distance) by using proximal cues that are probabilistically related to it (e.g., perceived size of an object, converging lines). For the system to respond successfully, Brunswik argued that cues should be utilized according to their ecological validity (see discussion in Hammond, 1966, p. 33), and that this concept is best measured by correlational statistics. Thus, ecological validity was defined as the correlation between a proximal cue and a distal criterion (Brunswik, 1952).

Unlike Hasher and Zacks's theory, Brunswikian theories of human judgment (e.g., Gigerenzer et al., 1991; Hammond et al., 1975) assume that the criterion—for instance, the frequency of sunny days in Rome in May—will typically *not* be directly retrieved from memory. Instead, it will be inferred based on proximal cues—for instance, the fact that Rome is located in southern Europe. Nevertheless, the Brunswikian research shares an interesting link with that of Hasher and Zacks (1984): While the latter assumes and provides evidence that people store accurate records of event frequencies, the former assumes and provides evidence that people keep fairly accurate records of ecological cue validities (e.g., Arkes & Hammond, 1986; Brehmer & Joyce, 1988). Learning cue validities, however, requires the ability to register event frequencies and their co-occurrences

accurately, except when knowledge of the validities is evolutionarily built in (e.g., in depth perception).

Inference by Subjective Cues: Availability

In a classic study by Tversky and Kahneman (1973), people had to judge whether each of five consonants (K, L, N, R, V) appears more frequently in the first or the third position in English words. Although all five consonants are more frequent in the third position, two-thirds of the participants judged the first position to be more likely for a majority of the letters.[4]

Tversky and Kahneman (1973) proposed the *availability heuristic* as a mechanism of real-world quantitative estimation that can account for systematic biases in people's estimates. According to the availability explanation, assessments of frequency (or probability) are based on the number of instances of the event that "could be brought to mind" (p. 207). That is, its basic assumptions are that people draw a sample of the event in question (e.g., by retrieving words that have the letter "R" in the first and third position, respectively) or assess the ease with which such a sample could be drawn, and then use the sample statistics to estimate the criterion. However, sample parameters may systematically deviate from population parameters (e.g., if it is easier to retrieve words with a certain letter in the first than in the third position, the sample will not be representative of the population). In this way, use of the availability cue may lead to systematic biases. Because the ability of a sample to predict the criterion can only be evaluated with respect to the sample drawn by a specific person, the availability cue is subjective rather than ecological.

Since Tversky and Kahneman (1973) proposed availability and other heuristics as important mechanisms underlying judgments of (relative) frequency and probability, their findings and the proposed heuristics have stimulated a tremendous amount of research and have raised serious concerns about people's ability to estimate event frequencies and probabilities accurately. At this point, the operation of availability is "one of the

4. In discussing Tversky and Kahneman's study, Lopes and Oden (1991) observed that 12 of the 20 English consonants are more frequent in the first position than in the third position, possibly explaining their results. In contrast, if one assumes that people have experienced a representative sample of letters and their positional frequencies (e.g., during reading), then their mental models should be well adapted to a representative sample presented by the experimenter. Sedlmeier et al. (1998) gave participants a representative sample of consonants (i.e., some that are more and some that are less frequent in the second position) and vowels. In each of three studies, they found that the estimated relative frequencies in the first versus the second position closely agreed with the actual rank ordering, except for an overestimation of low and underestimation of high values. Neither of the two versions of the availability heuristic that Sedlmeier et al. tested was able to account for these results.

most widely shared assumptions in decision making as well as in social judgment research" (Schwartz et al., 1991, p. 195). For example, it has been suggested that availability may account for people's tendency to exaggerate the frequency of some specific causes of death such as tornadoes (Lichtenstein et al., 1978) and for their performance in estimating demographic parameters such as countries' population size (Brown & Siegler, 1992, 1993).

Paradoxical Assumptions and Contradictory Findings

Here is the puzzle. Hasher and Zacks (1984) argued that people encode occurrences of an event, store a count of its frequency, and when required to estimate its frequency, access this count. Tversky and Kahneman (1973), in contrast, seemed to assume that people do not keep a record of event frequencies but construct a sample of the event in question and then infer event frequencies from the ease with which the sample could be constructed. Hasher and Zacks (1984) concluded that their experiments "reliably and unequivocably [sic] demonstrate remarkable knowledge of the frequency of occurrence of all events so far tested" (p. 1373), whereas Tversky and Kahneman (1973) took their results as evidence that the use of the availability heuristic leads to "systematic biases" (p. 209).

These contradictory assumptions and findings have been reported side by side in scientific journals and textbooks, without much discussion about how each line of research qualifies the other's findings (for exceptions, see Ayton & Wright, 1994; Holyoak & Spellman, 1993; Williams & Durso, 1986).[5] Suppose one tried to resolve the conflict by assuming that the two accounts—accurate judgments based on memorized experienced frequencies and (in)accurate judgments based on subjective cues—apply to different situations: The former holds whenever humans have experienced and encoded events one by one before making judgments, and the latter holds whenever humans have not directly experienced the criterion and thus have to rely on (subjective) cues correlated with it to derive a judgment.

This resolution, however, cannot work. Tversky and Kahneman's experiments also included situations where participants actually experienced the events sequentially. In one study, for instance, participants were serially presented with names of well-known personalities of both sexes (e.g., Elizabeth Taylor), and one group was then asked to judge whether the list contained more names of men or women (Tversky & Kahneman, 1973). Another example is their classic study of positional letter frequen-

5. One reason why this conflict did not attract more attention may be that Hasher and Zacks (1984) seem to have downplayed it. In a footnote they wrote: "The conflict between our view and that of Tversky and Kahneman is more apparent than real" (p. 1383; see their arguments in their footnote 9).

cies, mentioned above, in which they asked participants to judge events they had previously experienced sequentially. Both studies illustrate that availability is also intended to apply to experienced events.

In our view, the conflicting findings about the accuracy of people's frequency judgments and the conflicting claims about the underlying mechanisms cannot be reconciled simply by running more experiments in which people's estimates are observed to be either correct or incorrect. Contexts that elicit both biased and unbiased estimates can no doubt be found. The more interesting issue is how we can make theoretical progress in modeling the cognitive processes underlying quantitative estimation. Toward this goal, we pose two interrelated questions that are pertinent to both Hasher and Zacks's and Tversky and Kahneman's approaches. First, what do humans need to count in order to meet their adaptive goals? Second, what is the structure of the environments in which quantification occurs, and what heuristics can exploit that structure?

What Needs Counting?

The world can be carved up into an infinite number of discrete events or objects. Which of them deserve monitoring? Hasher and Zacks (1984, p. 1373) did not explicitly address this question, but proposed that for the frequency of a stimulus to be encoded and stored, it must at a minimum be "attended" to. The notion of "attention" was not precisely explicated.[6] How plausible is such a domain-general encoding mechanism, that is, a mechanism constrained only insofar that it requires attention (or "conscious" attention, as later proposed by Zacks et al., 1986)?

Consider, for instance, the processing that might occur when we walk down the street engaged in an engrossing conversation. We are generally successful at avoiding collisions with objects and other people, thus indicating that we take note of their locations. But later, would we be able to judge the relative frequency of their locations in relation to us (e.g., how many objects to the right and how many to the left of us), or the relative frequency of men and women who were wearing hats? Why should we be able to make such judgments retrospectively if we did not consider them useful at the time? More generally, do we encode every event and keep track of its frequency of occurrence, just because we have experienced it?

This is a question that neither the British empiricists nor Hasher and Zacks (1984) seriously addressed. For instance, David Hartley (1749) suggested a domain-general physiological mechanism of frequency counting

6. However, we can exclude one possible definition: intentional monitoring. Hasher and Zacks (1984) argued that a stimulus can be automatically encoded even if it is not intentionally monitored, which implies that intentional monitoring was not part of their definition of attention.

designed in analogy to Newton's theory of vibrations (Daston, 1988, p. 203). According to this mechanism, repeated occurrences of an object create cerebral vibrations until "grooves of mental habit" are etched into the brain. Hartley's is a content-general mechanism, insofar as it does not put any constraints on the type of objects to be counted. One can also find modern "cognitive" relatives of Hartley's physiological mechanism that are similarly unconstrained. Take, for instance, MINERVA 2 (e.g., Hintzman, 1984, 1988), which has been used to model frequency judgments. This model keeps copies (in terms of memory traces) of all events we have experienced over a lifetime (although one may bring content-specific considerations in through the back door by way of learning parameters, as Hintzman, 1988, does).

Should we be able to judge the relative frequency of men and women wearing hats? Marcia Johnson and her colleagues (Johnson et al., 1989) suggested that this is unlikely. On the basis of a series of ingenious studies, they demonstrated that, for frequency judgments to reflect presentation frequency accurately, two conditions must be met: The exposure time must be ≥2 seconds, and processing must involve directing attention to the identity of objects as well as their spatial location. Although their findings imposed an initial constraint on the mechanism, it remains essentially unconstrained with respect to what is counted.

Brase, Cosmides, and Tooby (1998) have proposed a more stringent constraint. They argued that another way to restrict the counting mechanism is to consider the nature of what is counted; there are aspects of the world that one would not expect a human inference mechanism to count spontaneously. According to their account, individuated whole objects rather than arbitrarily parsed objects (i.e., random chunks, nonfunctional fragments, etc.) are the natural unit of analysis: Toddlers may spontaneously count teddy bears, but not teddy bears' ears (as long as they have not been broken off the parent object).

A variation on Brase et al.'s approach is to consider the adaptive value of what is counted: Keeping track of event frequencies is most likely to occur in domains where knowing frequency counts has a plausible adaptive value for the organism. It is easy to see the value of monitoring the frequencies of specific events in the domains of mating and foraging (e.g., among the Inujjuamiut). But can considerations of adaptive value help us to derive counterintuitive predictions in other domains? We think so. Despite Tversky and Kahneman's (1973) seemingly unsupportive results (in their letter study), there is good reason to predict that people can quantify the statistical structure of language because of its adaptive value.

What Is Adaptive About Knowing the Statistical Structure of Language?

In any specific language certain sound sequences are more likely to occur within some words than in others. For instance, consider the sound se-

quence "pretty baby": The transition probability from "pre" to "ty" is greater than that from "ty" to "ba." Thus we would be more likely to expect a word break between the latter two syllables. For babies acquiring language, keeping track of these transition probabilities may have an important function, because these probabilities help them to identify boundaries between words (a problem that continues to hamper attempts to build a computer that "understands" spoken language). Recent results reported by Saffran, Aslin, and Newport (1996) indicate that babies are indeed sensitive to such transition probabilities.

To test whether babies have access to this kind of statistical information, Saffran et al. tested infants' ability to distinguish between "words" and "part-words" (using nonsensical stimuli in both cases). The stimulus words included sound sequences such as "bidaku" and "padoti" and a sample of the speech stream is "bidakupadotigolabubidaku. . . . " The babies listened to a two-minute tape of a continuous speech stream consisting of three-syllable words repeated in random order. A synthesized woman's voice spoke the sound stream with no inflection or noticeable pauses between words, removing the word boundary cues contained in normal speech. The only possible cues were the relative frequencies of co-occurrence of syllable pairs, where relatively low relative frequencies signal word boundaries.

After listening to the speech stream, the infants heard four three-syllable test words one at a time. Two words were from the speech stream and two were part-words. The part-words consisted of the final syllable of a word and the first two syllables of another word. Thus, a part-word contained sounds that the infant had heard, but it did not correspond to a word. Infants would be able to recognize part-words as novel only if the words from the original speech stream were so familiar to them that new sequences crossing word boundaries (i.e., the part-words) would sound relatively unfamiliar. In fact, the infants did listen longer to part-words than to words, indicating that they found them more novel than the words.

This example illustrates the importance of asking what information is adaptive to encode, store, and quantify. With this question in mind, one can derive interesting and counterintuitive predictions, for instance, that language learners will learn the statistical structure of language quickly. We now turn to the second important question: What is the structure of the environments in which quantities need to be estimated?

The Importance of "Ecological Texture"

"Although errors of judgments are but a method by which some cognitive processes are studied, the method has become a significant part of the message" (Kahneman & Tversky, 1982, p. 124). This quotation illustrates Kahneman and Tversky's awareness that the heuristics-and-biases pro-

gram came to focus on humans' cognitive errors at the expense of their cognitive successes. In fact, their initial framing of the availability heuristic stressed an ecological perspective that was later largely abandoned. Of the availability heuristic Tversky and Kahneman (1973) wrote:

> Availability is an ecologically valid clue for the judgment of frequency because, in general, frequent events are easier to recall or imagine than infrequent ones. However, availability is also affected by various factors which are unrelated to actual frequency. If the availability heuristic is applied, then such factors will affect the perceived frequency of classes and the subjective probability of events. Consequently, the use of the availability heuristic leads to systematic biases. (p. 209)

Not only did Tversky and Kahneman (1973) conceptualize availability as an "ecologically valid clue" to frequency, but they also stressed that it exploits the structure of the environment in the sense that objectively frequent events have stronger representations because these are strengthened by event repetitions, and thus, ceteris paribus, are easier to recall than infrequent ones. In light of its beginnings, the availability heuristic could have been developed into a cognitive strategy that reflects the texture of the environment as well as the mind, but was not.

Several decades ago, Egon Brunswik (1957) already emphasized the importance of studying the fit between cognition and the environment: "If there is anything that still ails psychology in general, and the psychology of cognition specifically, it is the neglect of investigation of environmental or ecological texture in favor of that of the texture of organismic structures and processes. Both historically and systematically psychology has forgotten that it is a science of organism-environment relationships, and has become a science of the organism" (p. 6).

In what follows, we propose an estimation heuristic that differs from those identified in the heuristics-and-biases program (e.g., availability) in several ways. First, how it exploits a particular environmental structure is specified. Second, it has a precise stopping rule that terminates memory search. Finally, it is formalized such that we can simulate its behavior. For these reasons, it exhibits bounded rationality. Before we analyze the structure of a specific class of environments in which various quantities have to be estimated, let us consider what adaptive value estimating one such quantity—population size—might have. We speculate that estimation of population demographics may be a descendant of an evolutionarily important task, specifically, estimation of social group size.

Estimation: Using Ecological Cues in a J-Shaped World

Because humans have always lived in groups (e.g., families, clans, tribes), it is very likely that social environments played a major role in shaping the human mind. Until recently, this possibility has largely been over-

looked in research on human reasoning and decision making. Wang (1996a, 1996b), however, demonstrated how social cues can affect decision making in surprising ways. Using Tversky and Kahneman's (1981) famous Asian disease problem, he found preference reversals (often considered irrational because they violate the invariance axiom of expected utility theory) when the text indicated that the decision was to be made for a large group. When the text indicated that the decision would affect a smaller group, however, most participants favored the risky outcome in both the loss and the gain framing.

Wang's (1996a) finding suggests that humans are sensitive to group size when making decisions. One may speculate that this sensitivity rests on an evolved ability to estimate group sizes. In fact, the ability to estimate the size of social groups accurately might have been of value in a number of circumstances encountered by our evolutionary ancestors, for instance, when they had to make quick decisions about whether to threaten to fight over resources with other families, clans, or tribes. Humans' social structures have changed since the time when we lived in hunter-gatherer societies. Group size has been directly affected by the shift from nomadic bands to small agricultural and pastoral communities to large populations of many thousands of people whose economic and social center is the city (e.g., Reynolds, 1973). Interestingly, in samples of American and Chinese participants, Wang (1996a) found that decision making is sensitive to culturally specific features of social group structure. Evolutionary considerations aside, we assume that the estimation heuristic proposed here is adapted to modern group sizes. We now consider the statistical structure of the environment in which the heuristic operates.

Let us start to analyze the statistical structure of population demographics by considering the following question. What distribution results if one makes a scatterplot of people's performance on the following task? Name all the characters in Shakespeare's *Comedy of Errors*. If we plotted people's performance on this task (e.g., the number of people who can name no, one, two, three, etc. characters), we would probably find that many people would get a low score, and that only a few people can attain a high score. Thus, contrary to the typical assumption of educational researchers that knowledge, learning, and performance generally conform to a bell-shaped distribution across individuals, in which moderate values are most frequent, human performance is often best characterized by the "empirical law of the higher, the fewer" (Walberg et al., 1984, p. 90), or in other words, by positively skewed, *J-shaped* distributions (where the "J" is rotated clockwise by 90 degrees).[7]

7. These distributions are related to Zipf's law (Zipf, 1949), which is the observation that frequency of occurrence of some event (P) as a function of the rank (i) when the rank is determined by the frequency of occurrence, is a power-law function $P_i \sim 1/i^a$ with the exponent a close to unity. The most famous example of Zipf's law is the frequency of English words. Assume that "the," "to," and "of" are the

Athletic performance can also follow such J-shaped distributions. Take the final distribution of medals in the 1996 Summer Olympics in Atlanta as an example. A total of 197 nations competed for 842 medals in the Atlanta games. Figure 10-1 plots the total number of medals won (gold, silver, and bronze) by each nation, excluding those that won no medals. The average number of medals won was 4.3. At one extreme, the United States, Germany, and Russia won 101, 65, and 63 medals respectively; in other words, 1.5 percent of the participating nations (and 8.5% of the world population) won almost one-third of all medals. At the other extreme, 118 participating nations won no medals at all. Highly positively skewed distributions also characterize many processes and phenomena in biology (e.g., fluctuations in neural spikes plotted by amplitude), geography (e.g., earthquakes plotted by severity), psychology (e.g., distribution of memory traces plotted by the likelihood they are needed; Anderson & Schooler, 1991), and other fields.

Cities plotted by actual population also form J-shaped distributions. In any given region, there are a few large settlements and a large number of small settlements. Herbert Simon (1955b) argued that in the special case of city population size, such a distribution is expected if the population growth is due solely to the net excess of births over deaths, and if this net growth is proportional to the present population size. Urban growth models that use techniques originally developed to model clumping and motion of particles in liquids and gases also predict this city size distribution (Makse et al., 1995). Figure 10-1 also shows the populations of German cities with more than 100,000 inhabitants ranked by their size. This distribution reflects the empirical law of the higher, the fewer in three ways: the largest value (here Berlin) is an extreme outlier; the mean (309,000), which is strongly influenced by such extreme observations, is much higher than the median (180,000); and the standard deviation (428,000) is large relative to the mean.

To what extent is it plausible to assume that people actually know about the J shape of distributions such as that of German cities? We asked 74 German participants to estimate the number of German cities in 25 size categories (100,000–199,999; 200,000–299,999; etc.). Figure 10-2 shows the distribution of their mean frequency judgments in comparison with the actual frequency distribution. (Note that compared with figure 10-1 the axes are reversed.) Although participants underestimated the relative number of cities in the smallest category (100,000–199,999), the results

---

three most frequent words (i.e., receive ranks 1, 2, and 3); then, if the number of occurrences is plotted as the function of the rank, the form is a power-law function with exponent close to 1. There are several variants of Zipf's law, such as Pareto's law, which essentially form J-shaped distributions. More generally, Grüneis et al. (1989) proved that J-shaped distributions belong to a class of distributions that can be modeled in terms of an adjoint Poisson process.

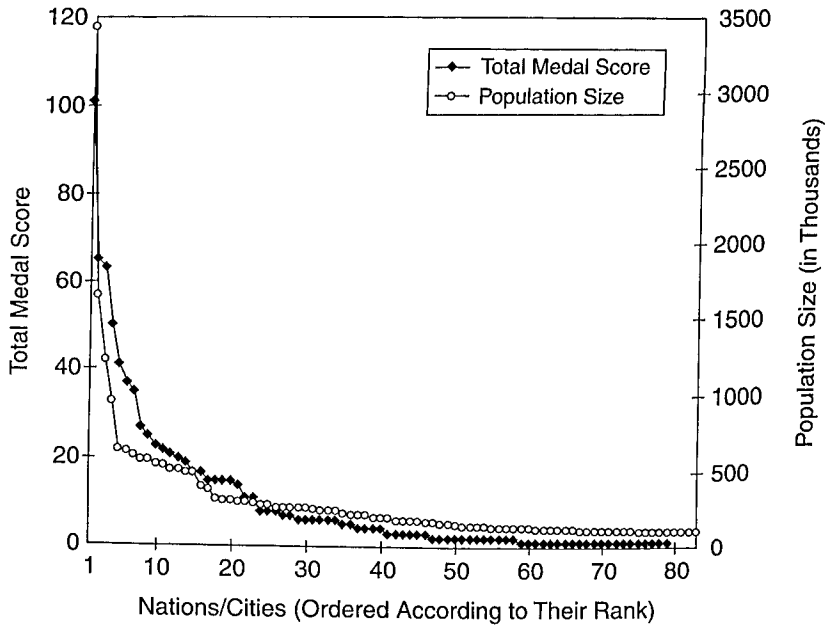


Figure 10-1: Distribution of medals won per nation at the 1996 Summer Olympics in Atlanta, and of the population size of the 83 largest German cities (*Fischer Welt Almanach*, 1993).

indicate that they were well aware of the skewness. Now that we have established that people have an intuition about the higher, the fewer characteristic of the German city size distribution, we turn to the next question: How might a heuristic exploit this J-shaped ecological structure so as to reduce the computational effort needed to make an estimate?

Fast and Frugal Estimation: The QuickEst Heuristic

Let us start by considering a technical problem, namely, sorting pieces of coal according to size. One way to sort them is to use a conveyor belt that carries the coal pieces across increasingly coarse sieves. The belt is designed so that first small pieces fall through the "small" sieve into the crusher below, then medium-sized pieces fall through the "medium" sieve, and so on. Pieces that make it across all the sieves are dumped into a catchall container. Let us assume that the sizes of the coal pieces follow a J-shaped distribution, that is, most pieces are small and only a few pieces are (very) large. The conveyor belt's design minimizes the time required for the sorting process by exploiting this fact, sorting out the large number of small pieces first, then the fewer larger ones, and finally the

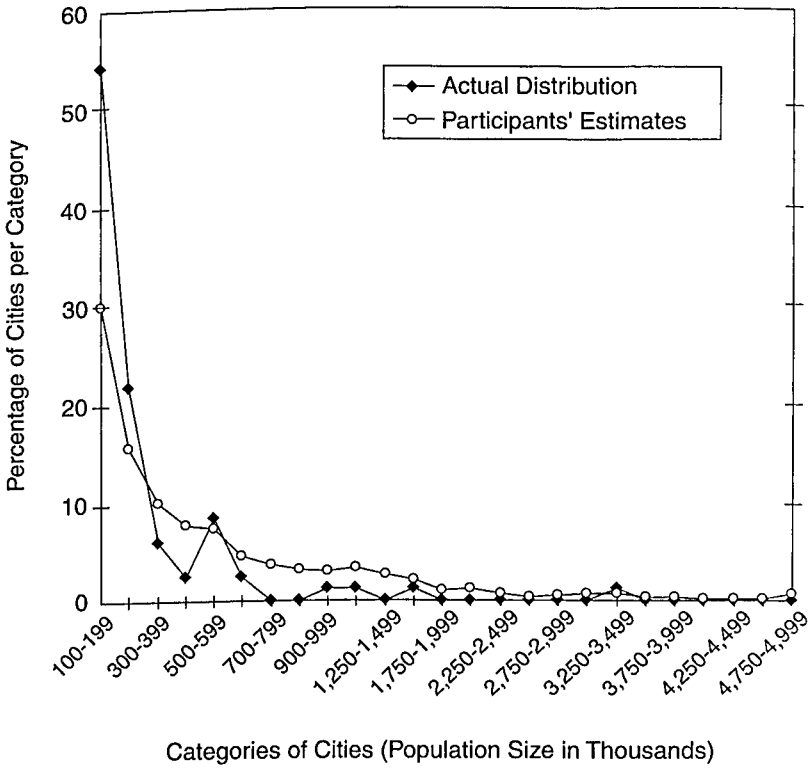


Figure 10-2: Percentage of German cities in 25 size categories, along with estimates made by participants (percentage values derived from frequency estimates).

very few largest ones. Figure 10-3 illustrates the design features of such a conveyor belt. We now propose an estimation heuristic, the Quick Estimation heuristic (QuickEst), which exploits the J-shaped distribution in a way similar to the conveyor belt for sorting coal.

QuickEst's Design Properties

QuickEst's policy is to use environmental structure to make estimates for the most common objects (e.g., in the cities environment, the smallest cities) as quickly as possible. What design features of the heuristic enable it to implement this policy?

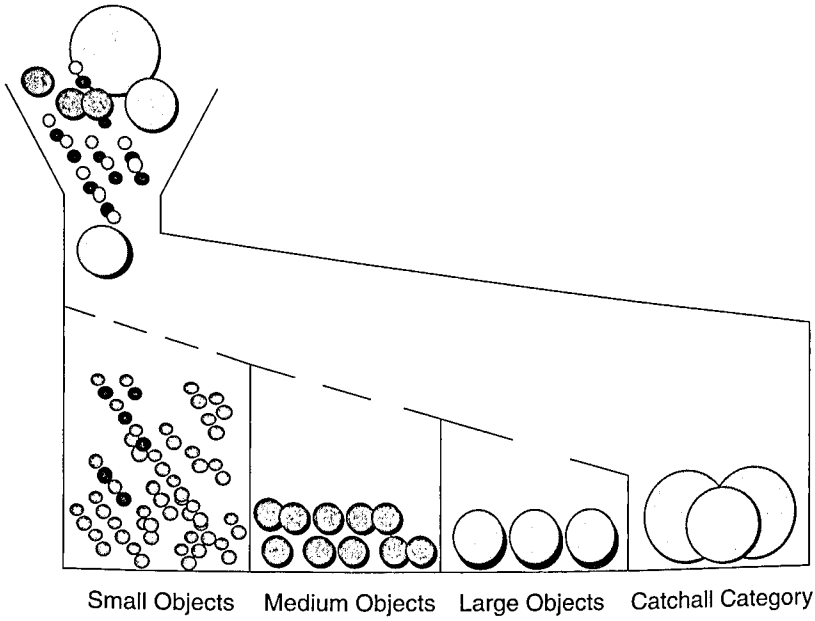


Figure 10-3: Illustration of a conveyor belt that sorts pieces of coal according to their size. (Although this is a fictitious example, its design resembles that of actual conveyor belts advertised at the Web sites of various manufacturers.)

How Are the Cues Ranked? When a person is asked to estimate the population of a city, the fact that it is a state capital may come to mind as a potential ecological cue. Cities that are state capitals (e.g., Munich, the capital of the state Bavaria) are likely to have larger populations than cities that are not state capitals, the major exceptions to this rule being in the United States. For any binary cue i, one can calculate the average size of cities that have this feature (s_i^+, e.g., the average size of all the German cities that are state capitals) and the average size of those cities that do not have this feature (s_i^-). Note that for the purpose of the simulations, we calculated s_i^- (s_i^+) from the actual sizes of the German cities that do not (or do) have the property. The input for this calculation, however, need not be the actual values, but could instead be imprecise subjective values.

Because positive cue values by definition indicate larger cities, s_i^- is smaller than s_i^+. For this reason, cues are ranked in the QuickEst heuristic according to s^-, with the smallest s^- first. This design follows the coal-sorting analogy, insofar as the cues (sieves) are ranked according to their coarseness, with the smallest cue first. For this ranking, the heuristic does

not need to know s^- exactly; it only needs to estimate a relative ranking of cues according to s^-.

When Is Search Stopped? Each cue asks for a property of a city, for instance, "Does the city have a university?" QuickEst has a simple stopping rule: Search is terminated when the first property is found that the city does not have (i.e., the response to the question is "no"). If a city has the property, then search continues, and its value on the cue with the next lowest s^- is retrieved from memory. This stopping rule has a *negative bias*, that is, a negative but not a positive value terminates search. This has an important consequence: As there are only a few cities with mainly positive cue values and many with mostly negative values, a stopping rule with such a negative bias generally enables the heuristic to stop earlier in the search and arrive at estimates quickly.[8]

Owing to its stopping rule, QuickEst's inference is based on the first property a city does not have. In contrast to computationally expensive strategies such as multiple regression, QuickEst does not integrate cue values. An important consequence of QuickEst's stopping rule is that the heuristic is noncompensatory. Further cue values (even if all of them are positive) do not change the estimate based on the first negative cue value encountered. By virtue of its simplicity, noncompensatory decision making avoids dealing with conflicting cues and the need to make trade-offs between cues.

How Coarse Are the Estimates? The estimate of QuickEst is the s^- of the first property a city does not have, rounded to the nearest *spontaneous number*. According to Albers (1997), spontaneous numbers are multiples of powers of 10 $\{a\ 10^i\colon a \in \{1, 1.5, 2, 3, 5, 7\}\}$, where i is a natural number. For instance, 300, 500, 700, and 1,000 are spontaneous numbers, but 900 is not. By building in spontaneous numbers, the heuristic takes into account two frequently observed properties of people's estimates. First, spontaneous numbers are related to what Albers (1997) described as number "prominence," that is, the phenomenon that in cultures that use the decimal system the powers of 10 "are the most prominent alternatives which have highest priority to be selected as responses, or terms by which given responses should be modified" (Albers, 1997, part I, p. 6). Second, spontaneous numbers relate to the phenomenon that, when asked for quantitative estimates (e.g., the price of a Porsche Carrera), people provide relatively coarse-grained estimates (e.g., \$70,000, i.e. 7×10^4 rather than \$75,342). This graininess of estimates, or crude levels of "relative exact-

8. For instance, in the reference class of all the German cities with more than 100,000 inhabitants and for the following *eight* ecological cues—soccer team, state capital, former East Germany, industrial belt, license plate, intercity train line, exposition site, and university (see chapter 4)—the German cities have on average about *six* (5.7) negative values.

ness" (Albers, 1997, part I, p. 12), reflects people's uncertainty about their judgments (see also Yaniv & Foster, 1995).[9]

The property that s^- is rounded to the nearest spontaneous number has two implications: First, for the numerical estimation the heuristic does not need to estimate s^-. It only needs to estimate which of two neighboring spontaneous numbers is nearer to s^-, and this spontaneous number is then given as the estimate.[10] Second, the heuristic's estimates can only achieve the precision and not exceed the graininess of spontaneous numbers.

How Can the Heuristic Deal With the Few Very Large Cities? The present stopping rule speeds up estimation by terminating search as soon as a property is found that the city in question does not have. Still, there are a handful of very large "outlier" cities that do have most properties. To avoid an unnecessarily time-consuming search for a possible property they do not have, QuickEst has a "catchall" category in reserve. That is, the heuristic stops adding more cues to its cue order as soon as most cities (out of those the heuristic "knows," i.e., the training set) have been sifted out. For our simulations, we assume that searching cues is stopped as soon as four-fifths of all the cities have already been sifted out by the heuristic. The remaining fifth of the cities are put into a catchall category and automatically assigned an estimate of s_j^+ (where cue j is the cue by which these largest cities were "caught" last) rounded to the nearest spontaneous number.

How Is QuickEst Ecologically Rational? QuickEst exploits the characteristics of the city population domain in two ways. First, its stopping rule—stop when the first negative cue value is found—limits the search process effectively in an environment in which negative cue values predominate. Second, its rank ordering of cues according to s^-, with the smallest s^- first, gives QuickEst a bias to estimate any given city as relatively small. This is appropriate for objects that fall in J-shaped distributions, in which most

9. Because there are more of them in the range of small digits (1, 1.5, 2, 3) than in the range of large digits (5, 7), spontaneous numbers also seem to be predicated on Benford's law. Benford's law (1938; Raimi, 1976) states that if numerical data (e.g., atomic weights) are classified according to the first significant digit, the nine classes that result usually differ in size. Whereas in a randomly generated data set, each number would be the first significant digit with frequency 1/9, in many real-world data sets, this frequency is approximately equal to $\log_{10}(p+1)/p$. Thus, the digit "1" is first about 30% of the time, "2" somewhat less often, and so on, with "9" occurring as the first digit less than 5% of the time. Consistent with Benford's law, 57% of German cities with more than 100,000 inhabitants begin with "1," whereas only 1.2% begin with "9."

10. Suppose that s^- lies in the interval between the spontaneous numbers 300,000 and 500,000. To decide whether s^- is to be rounded up or down, the heuristic only needs to know whether s^- belongs to the right or to the left of the interval's midpoint (i.e., 400,000). This only requires a choice (i.e., is s^- larger or smaller than 400,000).

objects have small values on the criterion, and only a few objects have (very) large values. In addition to being ecologically rational, QuickEst is psychologically plausible in that it provides estimates with the precision and graininess of spontaneous numbers.

Illustration

An American colleague of ours, Valerie, knows the approximate population size of five German cities from previous trips to Germany (Munich, 1,000,000; Frankfurt, 700,000; Nuremberg, 500,000; Bonn, 300,000; and Heidelberg, 150,000). Valerie also knows the cities' values on three cues (exposition site, state capital, and university). Given her limited knowledge about the reference class, German cities, how accurately could she infer the size of, for instance, Leverkusen? To answer this question, we first describe how QuickEst, as a model for Valerie's inferences, learns its parameters.

Training QuickEst ranks cues according to the average population size of cities that have negative values ($\bar{s}$). Given Valerie's knowledge, the cue with the smallest $\bar{s}$ is "exposition site," which provides the estimate 200,000.[11] The next cue is "state capital," which yields the estimate 500,000. Based on these two cues, the heuristic can sift out most of the cities Valerie knows: four out the five (i.e., 80%) have a negative value on at least one of these two cues. Thus, the only city that has positive values on the exposition site and state capital cues, Munich, is put into the catchall category. The estimate for this category is derived from the last cue in which Munich was "caught," here the state capital cue. The estimated size is 1,000,000 (which simply equals the size of Munich).

In sum, given Valerie's knowledge of German cities, the realization of the QuickEst heuristic includes two of the three cues she knows (exposition site and state capital), and a catchall category. This design allows QuickEst to derive one of three unique estimates for any given city in the reference class: 200,000, 500,000, and 1,000,000 inhabitants. How well does this realization of QuickEst perform when applied to new cities, for instance, Leverkusen and Hamburg?

Estimation To estimate the size of Leverkusen, QuickEst first retrieves that city's values on the exposition site cue. Because it does not have an exposition site, search is stopped and Leverkusen is estimated to have a population of 200,000, close to the 160,000 inhabitants it actually has. To derive an estimate for Hamburg, QuickEst looks up its value for the expo-

11. This figure is calculated as follows: Two of the five cities Valerie knows, Heidelberg and Bonn, do not have an exposition site. That is, $\bar{s}_{expo}$ equals the average size of Heidelberg and Bonn (225,000) rounded to the nearest spontaneous number (200,000).

sition site cue; as the value is positive, it then retrieves the value for the state capital cue, which is also positive. As a result, Hamburg ends up in the catchall category and is estimated to have a population of 1,000,000, which is not very close to the 1,650,000 inhabitants it actually has.

How good—or bad—is this performance, and how frugal is QuickEst in comparison with other heuristics?

Test of Performance: Environment and Competitors

To test QuickEst's performance more generally, we computed its estimates for the real-world environment of German cities with more than 100,000 inhabitants. After its reunification in 1990, Germany had 83 such cities. All of these cities (except Berlin) and their values on eight ecological cues to population size (the same cues as were used in chapter 4, except the national capital cue) were included in the test. (Berlin was excluded because it is an outlier and an error in estimating its population dwarfs errors of proportionally comparable size.) To evaluate the performance of QuickEst, we compared it with two competitors that demand considerably more computation and/or knowledge: multiple regression and an estimation tree (for quantification of the heuristics' complexity, see chapter 8).

Multiple regression is a demanding benchmark insofar as it calculates least-squares minimizing weights that reflect the correlations between cues and criterion, and the covariances between cues. Multiple regression has been proposed as both a descriptive and a prescriptive cognitive model, although its descriptive status is debated, given the complex calculation it assumes (for references on this issue, see chapter 4).

The second benchmark is an estimation tree (for more on tree-based procedures, see Breiman et al., 1993). With the aid of a computationally expensive Bayesian search process (e.g., chapter 8; Chipman et al., 1998), this tree was identified as one with a high probability of good performance.[12] It collapses cities with the same cue profile—that is, the same cue value on each of the eight ecological cues—into a class. The estimated size for each city equals the average size of all cities in that class. (The estimate for a city with a unique cue profile is just its actual size.) As long as the test set and training set are identical, this algorithm is optimal, and is equivalent to the exemplar-based algorithm model proposed by Persson (1996).[13] When the test set and training set are not identical the tree will

12. The Bayesian search was limited to the subset of trees that classified each new profile in the interval whose boundaries are defined by the cue profiles of known cities.

13. The optimal solution is to memorize all cue profiles and collapse cities with the same profile into the same size category. In statistics, this optimal solution is known as true regression and approximates the profile memorization method for optimal performance in choice tasks (see chapters 6 and 8).

encounter new cities with possibly new cue profiles. If a new city matches an old cue profile, its estimated size is the average size of those cities (in the training set) with that profile. If a new city has a new cue profile, then this profile is matched to the profile most similar to it. How is this done?

First, the cues are ordered within each profile according to their validity, with the one highest in validity first (for more on cue validity, see chapter 6). Second, the cue profiles are ordered lexicographically such that those with a positive value on the most valid cue are ranked first. Profiles that match on the first cue are then ordered according to their value on the second most valid cue, and so on. New cue profiles are filed with the lexicographically ordered old profiles according to the same logic. As an estimate of the size of a city with a new profile, the estimation tree takes the average size of those cities whose profile is above the new one in the lexicographical order. The estimation tree is an exemplar-based model that keeps track of all exemplars presented during learning as well as their cue values and sizes. Thus, when the training set is large, it requires vast memory resources (for the pros and cons of exemplar-based models, see Nosofsky et al., 1994).

We simulated population estimates, assuming varying degrees of knowledge about this environment. We tested a total of 10 sizes of training sets, in which 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 percent of the cities (and their respective sizes) were known. In the training phase, the three strategies—QuickEst, multiple regression, estimation tree—learned a model (or parameters) of the data (i.e., cities and their cue values; weights, s_j^+, s_j^-, etc.). To obtain reliable parameters, 1,000 random samples were drawn for each training set. For example, we drew 1,000 samples of 41 cities (50% training set) randomly from the reference class of 82 cities.

In the test phase, we applied the strategies to the complete reference class (i.e., test set, which includes the training set). The strategies' task was to estimate the populations of all the cities (assuming that the cities' values on the cues were known). To make the simulation psychologically more plausible, we assumed that the probability that a city belonged to the training set was proportional to its size. This assumption captures the fact that people are more likely to know about larger cities than smaller ones.

How Frugal Is QuickEst?

QuickEst is designed to make estimates quickly. How many cues must the heuristic consider before search is terminated? Figure 10-4 shows the number of cues that had to be retrieved by each strategy for various sizes of training sets. On average, QuickEst considers 2.3 cues per estimate—a figure that remains relatively stable across training sets. In contrast, multiple regression always uses all eight available cues. The estimation tree uses more and more cues as the size of the training set increases—across

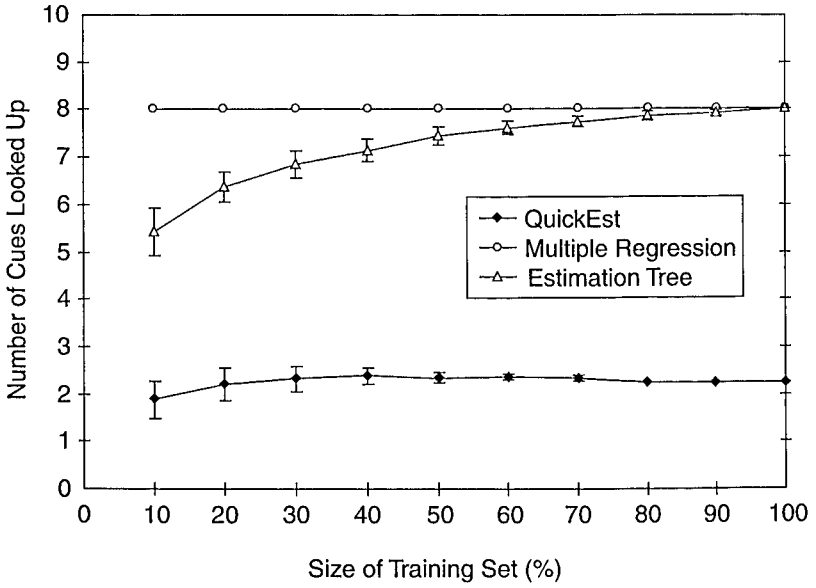


Figure 10-4: Number of cues looked up by QuickEst, multiple regression, and by the estimation tree as a function of size of training set. Vertical lines represent standard deviations.

all training sets, it uses an average of 7.2 cues. Thus, QuickEst bases its estimates on about 29% and 32% of the information used by multiple regression and the estimation tree, respectively.

How Accurate Is QuickEst?

How accurate is QuickEst, which involves simple averaging and rounding, compared with multiple regression, which involves complex calculations? We compared the three strategies' performance using two different measures of accuracy. First, we used the most common measure of estimation accuracy, according to Brown and Siegler (1993), that is, the (mean) absolute error (i.e., absolute deviation between actual and estimated size). Second, for the $(82 \times 81)/2$ city pairs in the complete set of paired comparisons, we simulated choices ("Which of the two cities is larger?") based on the estimates generated, and then calculated the proportion of correct inferences drawn.

Absolute Error

What price does QuickEst pay, in terms of absolute error, for considering only a few cues? Figure 10-5 shows the absolute error as a function of the

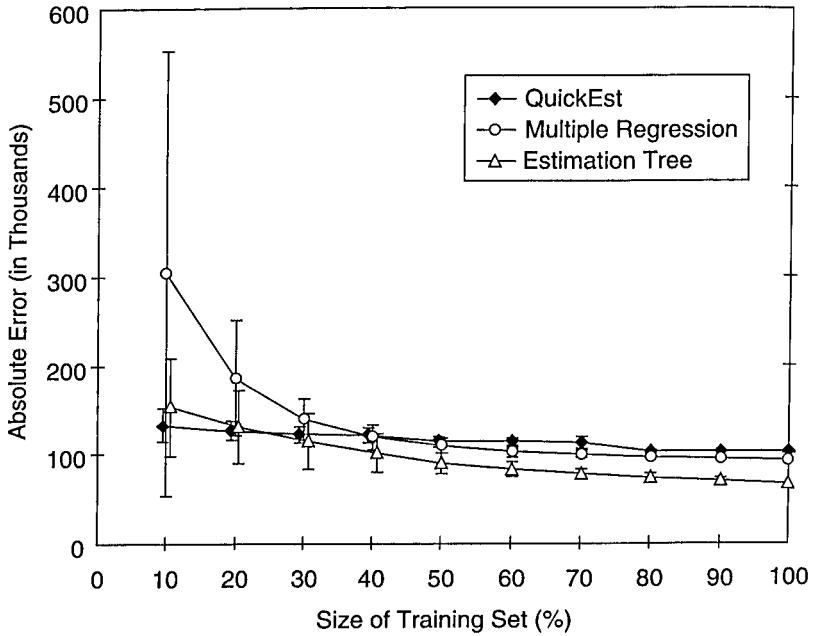


Figure 10-5: Mean absolute error (i.e., absolute deviation between predicted and actual size) as a function of size of training set. Vertical lines represent standard deviations. Note that some of the points have been offset slightly in the horizontal dimension to make the error bars easier to distinguish, but they correspond to identical training set sizes.

amount of learning (i.e., sizes of the training set). The 10% training set exemplifies a situation where knowledge is scarce (which is likely to be the rule rather than the exception in most domains). For this set, QuickEst's estimates are incorrect by an average of about 132,000 inhabitants (about half the size of the average German city in the simulated environment), compared with 303,000 for multiple regression, and 153,000 for the estimation tree. That is, under the psychologically relevant circumstances of scarce knowledge, QuickEst outperforms multiple regression clearly and the estimation tree by a small margin.

How does performance change as a function of learning (i.e., more cities known)? When 50% of the cities are known, for example, QuickEst and multiple regression perform about equally well, and the estimation tree outperforms both by a small margin. When the strategies have complete knowledge (all cities are known), multiple regression outperforms QuickEst by a relatively small margin—their respective absolute errors are about 93,000 and 103,000—and the estimation tree outperforms both competitors (absolute error is about 65,000, which equals the optimal performance, see footnote 13). That is, under the psychologically rather unlikely

circumstances of complete knowledge, QuickEst falls only slightly below the performance of multiple regression but is clearly outperformed by the estimation tree. (Even when multiple regression uses only those cues whose weights are significantly different from zero—7.3 on average instead of 8—its absolute error improves so slightly that the difference could hardly be seen if plotted in figure 10-5, except for the 10% training set.)

This result is similar to that reported by Chater et al. (1997). They tested the fast and frugal choice heuristic Take The Best (chapter 4), of which QuickEst is a relative, against four computationally expensive strategies, including neural networks and exemplar models. The task was to determine which of two German cities had the larger population size. Chater et al. found that when the training set was less than 40% of the test set, Take The Best outperformed all other competitors. Only when the training set grew beyond 40% did the competitors' performance increase above that of Take The Best.

Where does QuickEst make substantial errors? Figure 10-6 shows the deviations between actual and estimated size (in the 100% training set) for each strategy as a function of population size. Each heuristic has a distinct error pattern. Whereas QuickEst estimates the sizes of the many small cities quite accurately, it makes substantial errors on the few large

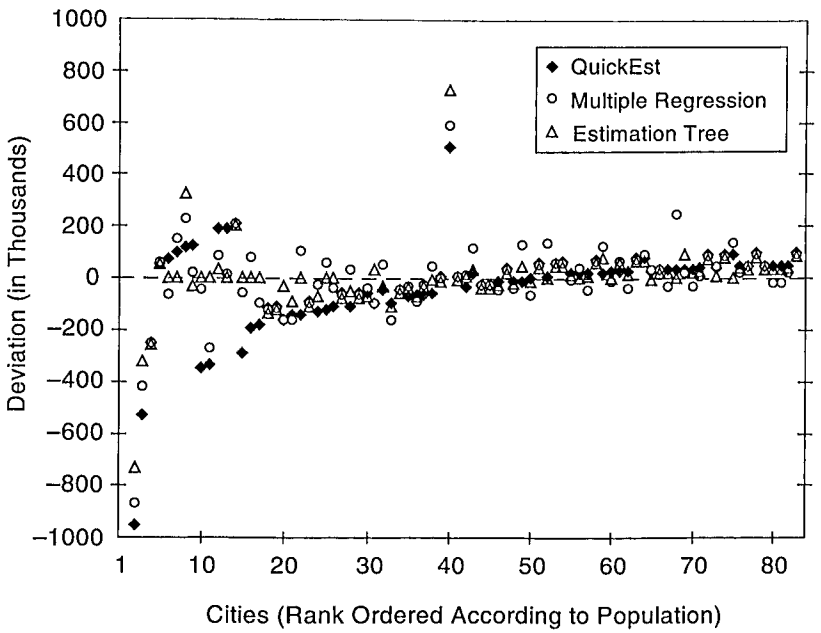


Figure 10-6: Deviation between actual and estimated size (in the 100% training set) for the three estimation methods on all cities, rank ordered according to population size.

cities because it puts them in its catchall category. Multiple regression, in contrast, makes substantial errors along the whole range of population size. The estimation tree makes relatively small errors for both small and large cities.

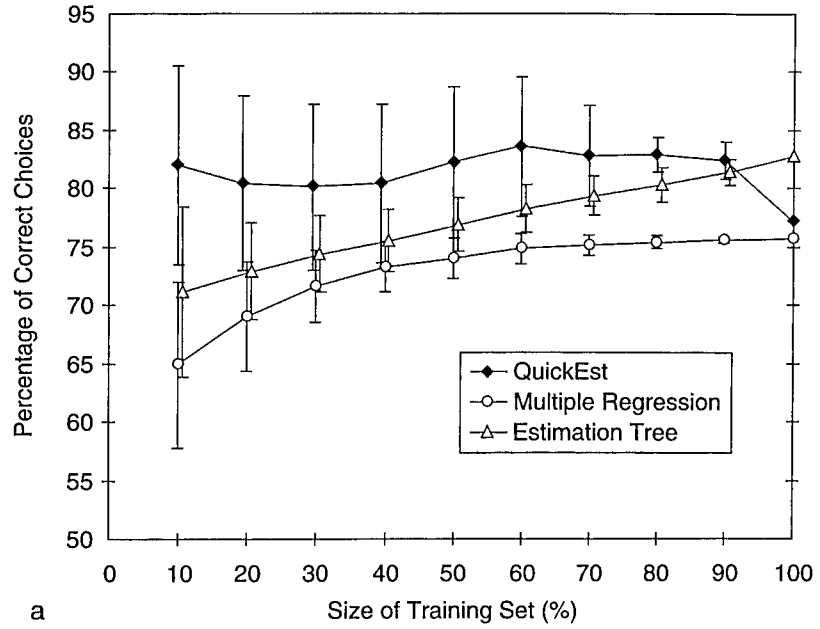
Another aspect of figure 10-6 deserves attention. More than the estimates made by the estimation tree and multiple regression, QuickEst's estimates are regressed toward the mean: On average, it underestimates the size of large cities and overestimates the size of small cities. Such a regression effect is typical in human quantitative estimation (e.g., Attneave, 1953; Lichtenstein et al., 1978; Sedlmeier et al., 1998; Varey et al., 1990). In figure 10-6, the overestimation of small city sizes appears miniscule compared to the underestimation of large city sizes. However, if the deviations between predicted and actual size are divided by actual size, then the regression effect for small cities is larger than for large cities. In the 100% training set, the median regression across all cities is 56%, 45%, and 23% for QuickEst, multiple regression, and the estimation tree, respectively (we applied the analysis described in Sedlmeier et al., 1998, footnote 1). Thus, QuickEst comes closest to showing the regression of about 70% observed in people's estimates in other tasks (Sedlmeier et al., 1998).

QuickEst uses only spontaneous numbers as estimates. What price will multiple regression pay if it has to work with the same psychological constraint? Recall that under complete knowledge (i.e., when all cities are known), multiple regression outperformed QuickEst (absolute errors of 93,000 vs. 103,000). If multiple regression also rounds its estimates to the nearest spontaneous number, however, it performs worse than QuickEst (absolute errors of 114,000 vs. 103,000).

To summarize, although the QuickEst heuristic involves only about a third of the information available to its competitors and fewer complex calculations than multiple regression, it outperforms multiple regression and the estimation tree when knowledge is scarce. In addition, QuickEst's performance is relatively stable across different amounts of learning: The absolute error is only 1.3 times higher for the 10% training set than for the complete knowledge case. In contrast, the absolute errors of multiple regression and the estimation tree in the 10% training set are 3.3 and 2.3 times higher than the absolute errors for complete knowledge, respectively. Only in the psychologically less plausible situation of abundant knowledge (i.e., 50% or more of the cities are known) is QuickEst (slightly) outperformed by its competitors.

Proportion of Correct Inferences

How many correct inferences do the heuristics make when comparing pairs of cities? Figure 10-7a shows the results for the proportion of correct inferences excluding cases of guessing (i.e., city pairs for which the heuristics chose randomly because the predicted sizes were identical), and



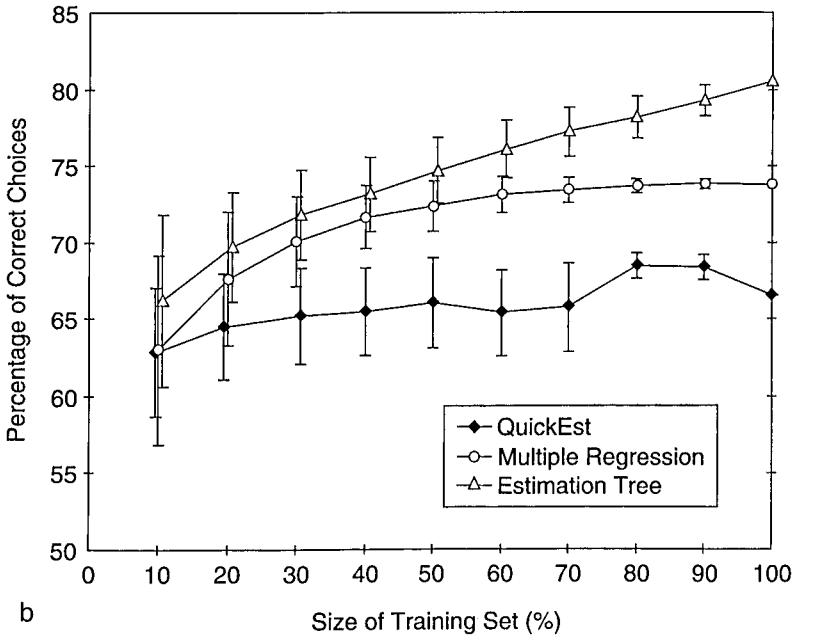


Figure 10-7: Percentage of correct city comparison inferences as a function of the size of training set, both excluding guessing (a) and including guessing (b). Vertical lines represent standard deviations. Note that some of the points have been offset slightly in the horizontal dimension to make the error bars easier to distinguish, but they correspond to identical training set sizes.

figure 10-7b shows the results including guesses. QuickEst's performance is excellent when it does not have to guess: Across all training sets, its proportion of correct inferences is 81%, whereas those of multiple regression and the estimation tree are 73% and 77%, respectively.

In cases in which the predicted sizes are identical, each of the strategies guesses randomly between the two cities, and thus, the proportion of correct inferences in such cases is expected to be 50%. Because this value is lower than the performance of the strategies without guessing, we can predict that overall performance decreases when guessing is included (see figure 10-7b). QuickEst suffers most because it falls back on guessing more because it has a smaller set of numerically distinct estimates available: Across all training sets, its proportion of correct inferences with guessing is 66%, whereas those of multiple regression and the estimation tree are 71% and 75%, respectively.

Conclusion

Let us conclude, as we began, with one of Robinson Crusoe's journal entries. Once Crusoe realized that his island was regularly visited by savages, he prepared himself for a possible confrontation with them. One early morning, he was surprised by

> seeing no less than five canoes all on shore together on my side of the island; and the people who belonged to them all landed, and out of my sight. The number of them broke all my measures; for seeing so many and knowing that they always came four, or six, or sometimes more, in a boat, I could not tell what to think of it, or how to take my measures, to attack twenty or thirty men single-handed; so I lay still in my castle, perplexed and discomforted. (Defoe, 1719/1980, p. 198)

For many evolutionarily important tasks, from choosing where to forage to deciding whether to fight, adaptive behavior hinges partly on organisms' ability to estimate quantities. Such decisions often have to be made quickly and on the basis of incomplete information. What structure of information in real-world environments can fast and frugal heuristics for estimation exploit to perform accurately? We presented a heuristic, QuickEst, that exploits a particular environmental structure, namely, J-shaped distributions. We demonstrated by simulation that where knowledge is scarce—as it typically is in natural decision-making settings (e.g., Klein, 1998)—the fast and frugal QuickEst outperforms or at least matches the performance of more expensive methods such as multiple regression and estimation trees. QuickEst is an ecologically rational strategy whose success highlights the importance of studying environmental structures.