How Forgetting Aids Heuristic Inference

Lael J. Schooler
Max Planck Institute for Human Development, Berlin

Ralph Hertwig
University of Basel, Switzerland

Schooler, L. J., Hertwig, R. (in press). How Forgetting Aids Heuristic Inference. *Psychological Review.*

Send correspondence to:

Lael Schooler
Max Planck Institute for Human Development
Center for Adaptive Behavior and Cognition
Lentzeallee 94
14195 Berlin
Germany
schooler@mpib-berlin.mpg.de

Abstract

A few theorists, ranging from William James to contemporary psychologists, have argued that forgetting should not be seen as a nuisance but as key for the proper working of human memory. In this paper, we propose the thesis that forgetting may in addition prove beneficial for inference heuristics that exploit mnemonic information such as recognition and retrieval fluency. To explore the mechanisms that link loss of information and heuristic performance, we bring together two research programs that take an ecological approach to studying cognition: fast and frugal heuristics (Gigerenzer, Todd, & the ABC Research Group, 1999) and ACT-R (Anderson & Lebiere, 1998). Using computer simulations, we show that forgetting boosts the accuracy of the recognition heuristic, which relies on systematic failures of recognition to infer which of two objects scores higher on a criterion value. Similarly, our simulations of the fluency heuristic, which arrives at the same inference on the basis of the speed with which objects are recognized, indicate that forgetting maintains the discriminability of recognition speeds. We conclude that the benefits of forgetting are not limited to generic advantages such as uncluttering memory from out-of-date information. Paradoxically, loss of mnemonic information allows these memory based heuristics to work more efficiently.

Introduction

In *The Mind of a Mnemonist*, Luria (1968) examined one of the most virtuoso memories ever documented. The possessor of this memory—S. V. Shereshevskii, to whom Luria referred as S.— reacted to the discovery of his extraordinary powers by quitting his job as a reporter and becoming a professional mnemonist. S.'s nearly perfect memory appeared to have "no distinct limits" (p. 11). Once, for instance, he memorized a long series of nonsense syllables that began "ma, va, na, sa, na, va, na, sa, na, ma, va" (Luria, 1968, p. 51). Eight years later, he recalled the whole series without making a single error or omission. This apparently infallible memory did not come without costs. S. complained, for example, that he had a poor memory for faces: "People's faces are constantly changing; it's the different shades of expression that confuse me and make it so hard to remember faces" (p. 64). "Unlike others, who tend to single out certain features by which to remember faces," Luria wrote, "S. saw faces as changing patterns . . . , much the same kind of impression a person would get, if he were sitting by a window watching the ebb and flow of the sea's waves" (p. 64). One way to interpret these observations is that cognitive processes such as generalizing, abstracting, and classifying different images of, for example, the same face requires ignoring the differences between them. In other words, crossing the "'accursed' threshold to a higher level of thought" (Luria, 1968, p. 133), which in Luria's view S. never did, may require the ability to forget.

Is forgetting a nuisance and a handicap, or is it essential to the proper functioning of memory and higher cognition? Much of the experimental research on memory has been dominated by questions of quantity, such as how much information is remembered and for how long (see Koriat, Goldsmith, & Pansky, 2000). From this perspective, forgetting is usually viewed as a regrettable loss of information. Some have suggested, however, that forgetting may be functional. One of the first to explore this possibility was James (1890), who wrote, "In the practical use of our intellect, forgetting is as important a function as recollecting" (p. 679). In his view, forgetting is the mental mechanism behind the selectivity of information processing, which in turn is "the very keel on which our mental ship is built."

A century later, Bjork and Bjork (1988) argued that forgetting prevents out-of-date information—say, old phone numbers or where one parked the car yesterday—from interfering with the recall of currently relevant information. Altmann and Gray (2002) make a similar point for the short-term goals that govern our behavior, such as keeping to the speed limit on a highway. From this perspective, forgetting prevents the retrieval of information that is likely obsolete. In fact, this is a function of forgetting that S. paradoxically had to do consciously. As a professional mnemonist, he committed thousands of words to memory. Learning to erase the images he associated with those words that he no longer needed to recall was an effortful, difficult process (Luria, 1968).

How and why forgetting might be functional has also been the focus of the extensive analysis conducted by Anderson and colleagues (Anderson & Milson, 1989; Anderson & Schooler, 1991; Anderson & Schooler, 2000; Schooler & Anderson, 1997). On the basis of their rational analysis of memory, they argued that much of memory performance, including forgetting, can be understood in terms of adaptation to the structure of the environment. The key assumptions of this rational analysis are that the memory system (1) meets the informational demands stemming from environmental stimuli by retrieving memory traces associated with the stimuli and (2) acts on the expectation that environmental stimuli tend to reoccur in predictable ways.

The rational analysis implies that memory performance reflects the patterns with which stimuli appear and reappear in the environment. To test this implication, Anderson and Schooler (1991) examined various environments that place informational demands on the memory system and found a strong correspondence between the regularities in the occurrence of information (e.g., a word's frequency and recency of occurrence) in these environments (e.g., conversation) and the classic learning and forgetting curves (as, for instance, described by Ebbinghaus, 1985/1964). In a conversational environment, for instance, Anderson and Schooler (1991) observed that the probability of hearing a particular word drops as the period of time since it was last used grows, much as recall of a given item decreases as the amount of time since it was last encountered increases. More generally, they argue that human memory essentially bets that as the recency and frequency with which a piece of information has been used decreases, the likelihood that it will be needed in the future also decreases. Because processing unneeded information is cognitively costly, the memory system is better off setting aside such little needed information by forgetting it.

In what follows, we extend the analysis of the effects of forgetting on memory performance to its effects on the performance of simple inference heuristics. To this end, we draw on the research program on fast and frugal heuristics (Gigerenzer, Todd, & the ABC Research Group, 1999) and the ACT-R research program (Anderson & Lebiere, 1998). Both programs share a strong ecological emphasis. The fast and frugal heuristics program examines simple strategies that exploit informational structures in the environment, enabling the mind to make surprisingly accurate decisions without much information or computation. The ACT-R research program strives to develop a coherent theory of cognition, specified to such a degree that phenomena from perceptual search to the learning of algebra can be modeled within the same framework. In particular, ACT-R offers a plausible model of memory that is tuned to the statistical structure of environmental events. This model of memory will be central to our implementation of the *recognition heuristic* (Goldstein & Gigerenzer, 2002) and the *fluency heuristic* (e.g., Jacoby & Dallas, 1981; Kelley & Jacoby, 1998), both of which depend on phenomenological assessments of memory retrieval. The former operates on knowledge about whether a stimulus can be recognized, while the latter relies on an assessment of the fluency, the speed, with which a stimulus is processed. By housing these memory-based heuristics in a cognitive architecture, we aim to provide precise definitions of heuristics and analyze whether and how loss of information—that is, forgetting—fosters the performance of these heuristics. We begin by describing the recognition heuristic, the fluency heuristic, and the ACT-R architecture.

How Recognition or Lack there of Enables Heuristic Inference: The Recognition Heuristic

Common sense suggests that ignorance stands in the way of good decision-making. The recognition heuristic belies this intuition. To see how the heuristic turns ignorance to its advantage, consider the simple situation in which one must select whichever of two objects is higher than the other with respect to some criterion (e.g., size or price). A contestant on a game show, for example, may have to make such decisions when faced with the question, "Which city has more inhabitants, San Diego or San Antonio?" How she makes this decision depends on the information available to her. If the only information on hand is whether she recognizes one of the cities and there is reason to suspect that recognition is positively correlated with city population, then she can do little better than rely on her (partial) ignorance. This kind of ignorance-based reasoning is embodied in the recognition heuristic (Goldstein & Gigerenzer, 1999, 2002), which for a two-alternative choice can be stated as follows:

> If one of two objects is recognized and the other is not, then infer that the recognized object has the higher value with respect to the criterion.

Partial ignorance may not sound like much for a decision maker to go on. Because lack of recognition knowledge is often systematic rather than random, however, failure to recognize something may be informative. The recognition heuristic exploits this information.

Empirical Support

To find out whether people use the recognition heuristic, Goldstein and Gigerenzer (1999, 2002) pursued several experimental approaches. In one approach, they presented University of Chicago students with pairs of cities randomly drawn from the 22 largest cities in the United States and pairs of cities randomly drawn from the 22 largest cities in Germany. The task was to infer which city in each pair had the larger population. The students at this American university made a median of 71% correct inferences in the American city set. When quizzed on the German city pairs, they made a median of 73% correct inferences. If one assumes that more knowledge leads to better performance, these results are counterintuitive: Years of living in the United States gave these students ample opportunity to learn facts about American cities that could be useful for inferring city populations, whereas they knew little or nothing about the German cities beyond recognizing about half of them. Why would they perform slightly better in the German city set? According to Goldstein and Gigerenzer (2002), the American students' meager knowledge about German cities is precisely what allowed them to infer that the cities they recognized were larger than the cities they did not recognize. The recognition heuristic was of no use to them when making judgments about American cities, because they recognized all those cities. Goldstein and Gigerenzer referred to this surprising phenomenon as the *less-is-more effect* and showed analytically that recognizing an intermediate number of objects in a set can yield the highest proportion of correct answers. All else being equal, recognizing more than this many objects decreases inferential accuracy.

The following example, adapted from Goldstein and Gigerenzer (2002), provides an intuitive illustration of how the recognition heuristic gives rise to the less-is-more effect. Suppose that three brothers have to take a test on the 20 largest German cities. The youngest brother has never heard of any of the cities, the middle brother has heard of 10 of them, and the eldest brother has heard of them all. The middle brother tends to know the names of the larger cities. In fact, the 10 cities he recognizes are larger than the 10 cities he does not in, say, 80 of the 100 possible pairs to which he can apply his recognition knowledge (i.e., the 100 pairs in which he recognizes one city and does not recognize the other). Thus, his *recognition validity* (i.e., the proportion of times that a recognized object has a higher value on the criterion than does an unrecognized object in a given sample) is .8. Both the middle and the eldest brothers have some knowledge of German cities aside from recognition. When they recognize both cities in a pair, they have a 60% chance of correctly choosing the larger one on the basis of this other knowledge, so their *knowledge validity* is .6.

Suppose the tester randomly draws pairs from the twenty largest German cities and asks the three brothers to decide which member of each pair has the larger population. Who will score highest? The youngest brother guesses the answer to every question and thus gets 50% correct. The eldest brother relies on his knowledge about the cities to answer every question and scores 60% correct. Neither the youngest brother nor the eldest brother can use the recognition heuristic, the former because he fails to recognize any of the cities and the latter because he recognizes them all. The only one with partial ignorance to exploit, the middle

brother makes 68% correct inferences, thus surpassing the inference accuracy of the eldest brother, because his recognition validity of .8 exceeds the elder brother's knowledge validity of .6.[1]

How Recognition Exploits Environmental Correlations

How can people learn the association between recognition and a criterion when the criterion is not accessible? Goldstein and Gigerenzer (2002) proposed that there are "mediators" in the environment that both reflect the criterion and are accessible to the decision maker. For example, although a person may not know the population size of a German city, its size may be reflected in how often it is mentioned in the person's environment. This frequency of mention, in turn, is correlated with how likely the person is to recognize the city's name. This chain of correlations would enable people to make inferences about a city's size on the basis of whether they can recognize it. To test the extent to which environmental frequencies can operate as mediators between city recognition and city population, Goldstein and Gigerenzer (2002) computed the correlations among three measures for each of the 83 German cities with more than 100,000 inhabitants: the actual population, the number of times the city was mentioned in the *Chicago Tribune* over a specific period, and the *recognition rate,* that is, the proportion of University of Chicago participants who recognized the city (see upper portion of Figure 1).

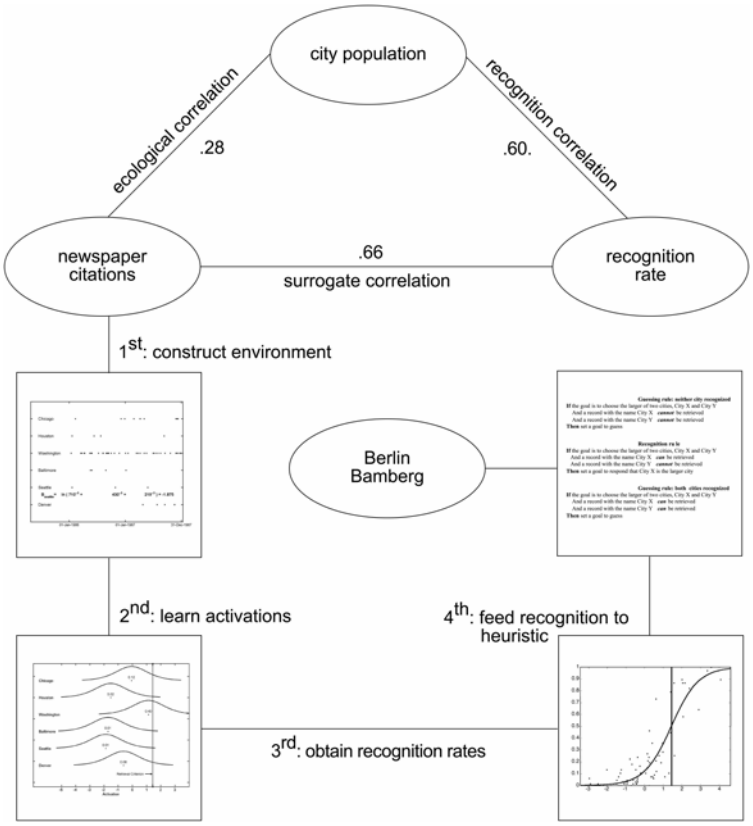

Figure 1. The triangle on top, adapted from Goldstein & Gigerenzer's (2002) Figure 7, shows the Pearson correlations between how often a German city was mentioned in the *Chicago Tribune*, its population, and its recognition rate. The lower portion outlines the steps in the simulation described in the text. 1) Environments are constructed based on word frequency information from the Chicago Tribune. 2) The base level activations the city name records were learned from the environment. 3) The model's recognition rates were obtained by fitting it to the student's recognition rates. 4) The recognition heuristic was applied to the German city task, drawing on the simulated recognition rates.

The *ecological correlation*, that is, the correlation between how often a city was mentioned in the *Chicago Tribune* and its population, was .82. Does newspaper coverage of a city correlate with the number of people who recognized it? Yes, the *surrogate correlation*, that is, the correlation between how often a city was mentioned and the number of people recognizing it, was .66. Finally, the correlation between the number of people recognizing a city and its population, known as the *average recognition validity*, was .60. In other words, the cities' recognition rates were more closely associated with how often they were mentioned than with their actual populations. Because recognition tracks environmental frequency more closely than it tracks population size in this context, Goldstein and Gigerenzer (2002) suggested that environmental frequency is the mediator between recognition and population size.

The recognition heuristic relies on ignorance that is partial and systematic. It works because lack of recognition knowledge about objects such as cities, colleges, sports teams, and companies traded on a stock market is often not random. For Goldstein and Gigerenzer (2002), a lack of recognition comes from never having encountered something, dividing the "world into the novel and previously experienced" (p. 77). If human recognition were so exquisitely sensitive to novelty that it treated only those objects and events as unrecognized that one has never seen, then experience would eventually render the recognition heuristic inapplicable (see Todd & Kirby, 2001). Like the ignorance that comes from lack of experience, forgetting may maintain or even boost inferential accuracy by making the old, novel again. For illustration, consider the oldest brother in the three-brother scenario. If he were able to forget some of the city names, he could take advantage of the recognition heuristic. The resulting changes in his performance would depend on which cities he no longer recognized. If his forgetting were random, he could not effectively exploit his "recovered" ignorance. If he tended to forget the names of smaller cities (of which he is likely to have heard about much less frequently), however, he could benefit from his ignorance. The recognition heuristic, however, may not be the only inference strategy that could benefit from forgetting.

How Retrieval Fluency Enables Heuristic Inference: The Fluency Heuristic

A key property of heuristics is that they are applicable under limited circumstances that, ideally, can be precisely defined. The recognition heuristic, for example, cannot be applied when both objects are either recognized or unrecognized. Thus, if a person's recognition rate is either very low or very high, she can rarely use the heuristic. When she does not recognize either object, use of the recognition heuristic gives way to, for instance, guessing. When she recognizes both objects, more knowledge-intensive strategies, such as the *Take The Best heuristic,* can be recruited (Gigerenzer, Todd, & the ABC research group, 1999). Take The Best sequentially searches for cues that are correlated with the criterion in the order of their predictive accuracy and chooses between the objects on the basis of the first cue found that discriminates between them (Gigerenzer & Goldstein, 1996).

Another, less knowledge intensive, inference strategy that can be applied to a two-alternative choice when both objects are recognized is the fluency heuristic (see, e.g., Jacoby & Brooks, 1984; Toth & Daniels, 2002; Whittlesea, 1993). It can be expressed as follows:

> If one of two objects is more fluently processed, then infer that this object has the higher value with respect to the criterion.

Like the recognition heuristic, the fluency heuristic relies on only one consideration to make a choice; in this case, the fluency with which the objects are processed when encountered. In numerous studies, processing fluency mediated by prior experience with a stimulus has been

shown to function as a cue in a range of judgments. For example, more fluent processing due to previous exposure can increase the perceived fame of nonfamous names (the false fame effect; Jacoby, Kelley, Brown & Jasechko, 1989) and the perceived truth of repeated assertions (the reiteration effect; Begg, Anas, & Farinacci, 1992; Hertwig, Gigerenzer, & Hoffrage, 1997).

As we show shortly, the ACT-R architecture offers the possibility of precisely defining fluency and how it depends on the past history of environmental exposure. As in the case of the recognition heuristic, the ACT-R architecture allows us to examine how forgetting may affect the fluency heuristic's accuracy. But unlike the recognition heuristic, the fluency heuristic seems to reflect the common intuition that more information is better (see Hertwig & Todd, 2003). To appreciate this, let us return to the oldest of the three brothers, who recognizes all the twenty largest American cities. If his history of exposure to the cities, mediated by this history's effect on fluency, were indicative of their population size, he may now be able to match or even outdo the performance of the middle brother. To figure out which brother will do best, one needs to know the two heuristics' *validities* (the percentage of correct inferences that each yields in cases where it is applicable) and *application* rates (to what proportion of choices can each heuristic be applied).

The fluency heuristic, in contrast to the recognition heuristic, does not exploit partial ignorance, but rather graded recognition. Could it also benefit from forgetting? This is one of the key questions that we address in our analysis. Specifically, we investigate the role of forgetting in memory-based heuristics by modeling the relation between environmental exposure and the information in memory on which heuristics such as recognition and fluency feed. To lay the necessary groundwork, we now provide a brief introduction to the ACT-R architecture and describe how we implemented the recognition and fluency heuristics within it.

A Brief Overview of ACT-R

ACT-R is a unified theory of cognition that accounts for a diverse set of phenomena ranging from subitizing (Peterson & Simon, 2000) to scientific discovery (Schunn & Anderson, 1998). A central distinction in ACT-R is between declarative knowledge ("knowing that") and procedural knowledge ("knowing how"). ACT-R models procedural knowledge with sets of production rules (i.e., if-then rules) whose conditions (the "if" part of the rule) are matched against the contents of declarative memory. The fundamental declarative representation in ACT-R is the chunk, which we refer to here as a *record* to highlight the parallels between memory retrieval and information retrieval in library science. If all the conditions of a production rule are met, then the rule fires, and the actions specified in the "then" part of the rule are carried out. These actions can include updating records, creating new records, setting goals, and initiating motor responses. For example, Table 1 shows a set of colloquially expressed production rules that implement the recognition heuristic.

Which of the rules in Table 1 will apply depends on whether records associated with City *X* and City *Y* can be retrieved. The overall performance of the recognition heuristic depends on how often each of these rules applies and how accurate the inferences based on each are when it does apply. Hereafter we refer to the complete set of rules in Table 1 as the recognition heuristic and to the second rule specifically as the *recognition rule*.

Table 1. The production set that implements the Recognition Heuristic

Guessing rule: neither city recognized
If the goal is to choose the larger of two cities, City X and City Y
And a record with the name City X ***cannot*** be retrieved
And a record with the name City Y ***cannot*** be retrieved
Then set a goal to guess

Recognition rule
If the goal is to choose the larger of two cities, City X and City Y
And a record with the name City X ***can*** be retrieved
And a record with the name City Y ***cannot*** be retrieved
Then set a goal to respond that City X is the larger city

Guessing rule: both cities recognized
If the goal is to choose the larger of two cities, City X and City Y
And a record with the name City X ***can*** be retrieved
And a record with the name City Y ***can*** be retrieved
Then set a goal to guess

In ACT-R, declarative memory and procedural memory interact through retrieval mechanisms that assume that certain events tend to reoccur in the environment at certain times. In essence, the records that the system retrieves at a given point can be seen as a bet about what will happen next in the environment. In this framework, human memory functions as an information retrieval system, and the elements of the current context constitute a query to long-term memory to which the memory system responds by retrieving the records that are most likely to be relevant.

Many word processors incorporate a timesaving feature that, like ACT-R, takes advantage of forgetting. When a user goes to open a document file, the program presents a "file buffer", a list of recently opened files from which the user can select. Whenever the desired file is included on the list, the user is spared the effort of searching through the file hierarchy. For this device to work efficiently, however, the word processor must provide users with the files they actually want. It does so by "forgetting" files that are considered unlikely to be needed on the basis of the assumption that the time since a file was last opened is negatively correlated with its likelihood of being needed now. Similarly, the declarative retrieval mechanism in ACT-R makes more recently retrieved memory records more accessible on the assumption that the probability that a record is needed now depends in part on how long ago it was last needed.

Conducting Search

ACT-R makes the assumptions that information in long-term memory is stored in discrete records and that retrieval entails searching through these records to find a record that achieves some processing goal of the system. The explanatory power of the approach depends on the system's estimates of the probability that each record in long-term memory is the one sought. In keeping with common usage in library science, we call this the *relevance probability*.[2] Any information retrieval system must strike a balance between the rate of recall, in this context the likelihood of finding the desired record (i.e., the proportion of hits), and the precision of recall, or the likelihood of retrieving irrelevant records (i.e., the proportion of false alarms). In ACT-R, this balance is achieved through a guided search process in which the records are

retrieved in order of their relevance probabilities, with the most promising records looked up first.

Stopping Search

At some point, the information retrieval system must terminate search for further records. If p is the relevance probability, C is the cost of attempting to match a memory record against a condition of a production rule, and G is the gain associated with successfully finding the target, then according to ACT-R the memory system should stop considering records when:

$$pG < C \tag{1}$$

In other words, the system stops searching for more memory records when the expected value (pG) of the next record is less than the cost of considering it. If the next record to be considered has a relevance probability of less than C/G, search will be stopped.

Activation Reflects Relevance

In ACT-R, the activation of a declarative memory record reflects its relevance probability. Specifically, the activation, A, equals the log odds (i.e., $\ln[p/(1-p)]$) that the record will be needed to achieve a processing goal (i.e., that it will match a condition of a production rule that fires). A record's activation is calculated by a combination of the base-level strength of the record, B_i, and the S_{ji} units of activation the record receives from each of the j elements of the current context:

$$A_i = B_i + \sum_j S_{ji} \tag{2}$$

A record's base-level strength is rooted in its environmental pattern of occurrence. Specifically, B_i is determined by how frequently and recently the record has been encountered in the past:

$$B = \ln\left(\sum_{j=1}^{n} t_j^{\,d}\right), \tag{3}$$

where the record has been encountered n times in the past and the j^{th} encounter occurred t_j time units in the past. Finally, d is a decay parameter that captures the amount of forgetting in declarative memory, thus determining how much information about an item's environmental frequency is retained in memory.[3] Typically, d is set to equal –.5, a value that has been shown to fit a wide range of behavioral data (Anderson & Lebiere, 1998).

Consider, for illustration, the occurrence of American city names in the front-page headlines of the *New York Times*. Each circle in Figure 2 indicates a day on which a particular city appeared in the front-page headlines between January 1, 1986, and December 31, 1987. Clearly, there are drastic differences between cities in their frequency of occurrence. The national capital, Washington, DC, was mentioned 37 times during that period, first on January 14, 1986, and last on December 26, 1987. Seattle, in contrast, was mentioned merely 3 times—710, 430, and 219 days before January 1, 1988. Figure 2 also shows how these quantities are used to determine base-level activation on January 1, 1988. For this calculation,

the parameter *d,* the decay rate, was set to –.5, and the resulting activation for Seattle, for example, is $\ln(710^{-.5} + 430^{-.5} + 219^{-.5}) = -1.87$.

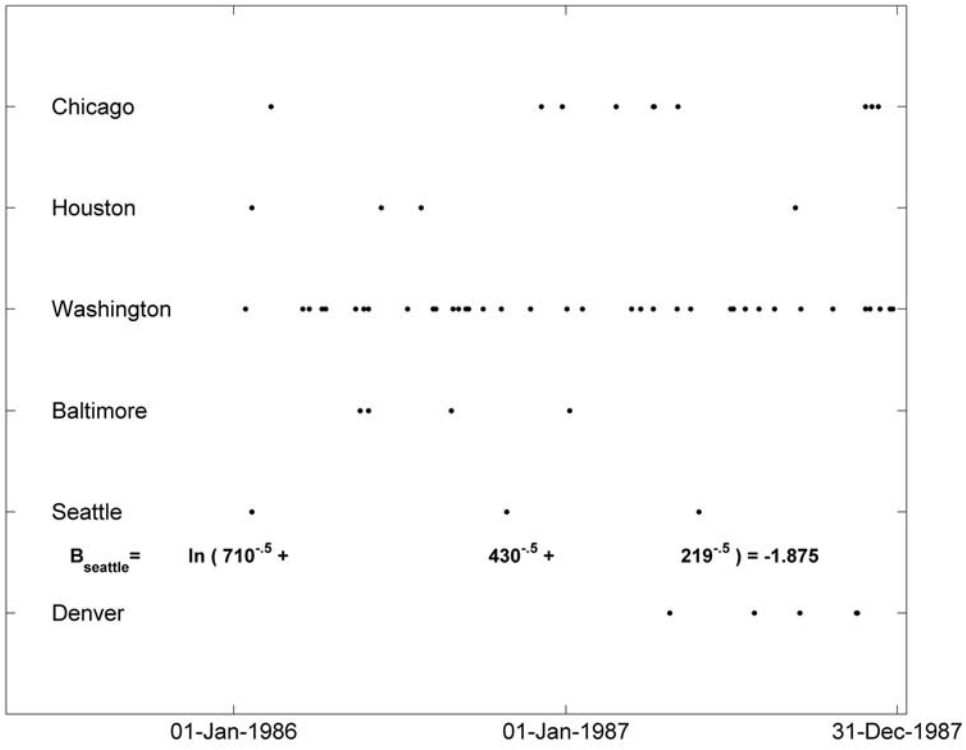

Figure 2. Number of days on which various city names were mentioned on the front page of the New York Times headlines.

Activation and Retrieval Probability

Whether a memory record's activation exceeds the retrieval criterion is determined by a noisy process. The sources of noise include momentary fluctuations in a record's estimated gain, estimated cost, and the activation it receives from the current constellation of context elements. These context elements could be part of our external environment, such as words on a sign, or internal, such as our mood or recently activated records. In the simulations reported below, we do not model the influence of contextual information, represented by the second term in Equation 2, in detail but rather assume that it contributes to the overall variance in activation. Because of its variability, the activation of a memory record is better represented by a distribution of activation values than by a single value, with *B* (see Equation 3) as the distribution's expected value. In ACT-R, activation is modeled as a logistic distribution, which approximates a normal distribution. Figure 3 shows these distributions around the expected value of the activation for the cities depicted in Figure 2.

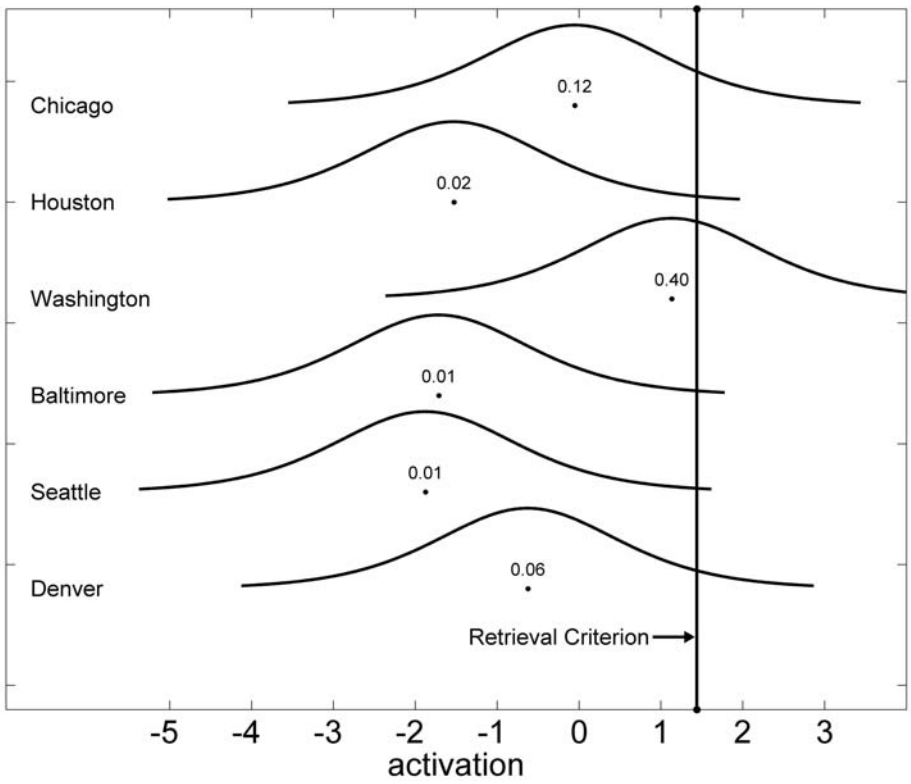

Figure 3. Activation distributions for the city name records that result from being mentioned in the *New York Times* headlines. The points correspond to the expected value of the distribution. The number at the center of each distribution shows the proportion of the distribution that lies to the right of the retrieval criterion τ (which to be consistent with subsequent simulations is set to 1.44). This proportion is the probability that the system will be able to retrieve the record and thereby recognize the city.

The probability that a record will be retrieved, that is, that its activation will exceed the retrieval criterion, can be expressed as a logistic function:

$$\text{probability of record retrieval} = \frac{1}{1+e^{-(A-\tau)/s}} \quad , \qquad (4)$$

where *s* captures momentary and permanent fluctuations in the activation level of record *i.* Parameter τ equals ln C/ (G–C), a stopping rule that is equivalent to the p< C/G-criterion from Equation 1 but transformed into the activation scale. The proportion of a record's activation distribution that is above the retrieval criterion, τ, gives the probability that the particular record will be retrieved. Retrieval of the memory record is crucial for our analysis of the recognition heuristic because we adopt Anderson, Bothell, Lebiere, and Matessa's (1998) assumption that retrieval of a record implies recognition of the associated word or, in this case, of the city name. The retrieval criterion τ is typically estimated by fitting models to data. The value of 1.44 used in Figure 3 is taken from the subsequent simulations. As Figure 3 shows, about one-twentieth of the activation distribution for Denver, for example, lies to the right of the retrieval criterion, corresponding to a 6% chance that Denver will be recognized.

In brief, the relevance probability of each memory record is reflected in its distribution of activation. Records are searched in order of their activation until either a record is found that satisfies the current condition or the activation of the next record to be considered is so low that it is not worth considering.[4]

Activation and Retrieval Time

In ACT-R, retrieval time is an exponential function of activation:

$$\text{retrieval time for a record} = Fe^{-A} \qquad (5)$$

where A is the activation for a particular record and F is a scale parameter. Anderson et al. (1998) found that values of F can be systematically estimated from the retrieval threshold, τ, using the equation $F=.348e^{\tau}$, so τ of 1.44 yields a value of 1.47 for F. Figure 4 plots Equation 5 for these parameter values, and the solid line represents the range of retrieval times that would be observed for activation values exceeding the retrieval threshold. As Figure 4 shows, the lower the activation, the more time it takes to retrieve a record. The open circle represents the retrieval time for a memory record whose activation falls just above τ. A memory record with activation below this point will fall short of the retrieval criterion and, because records with such low activations are unlikely to achieve processing goals, will fail to be retrieved at all. As Figure 4 reveals, subsequent increases in activation lead to diminishing reductions in retrieval time, a property that, as we will see shortly, is crucial to understanding how forgetting impacts the fluency heuristic.

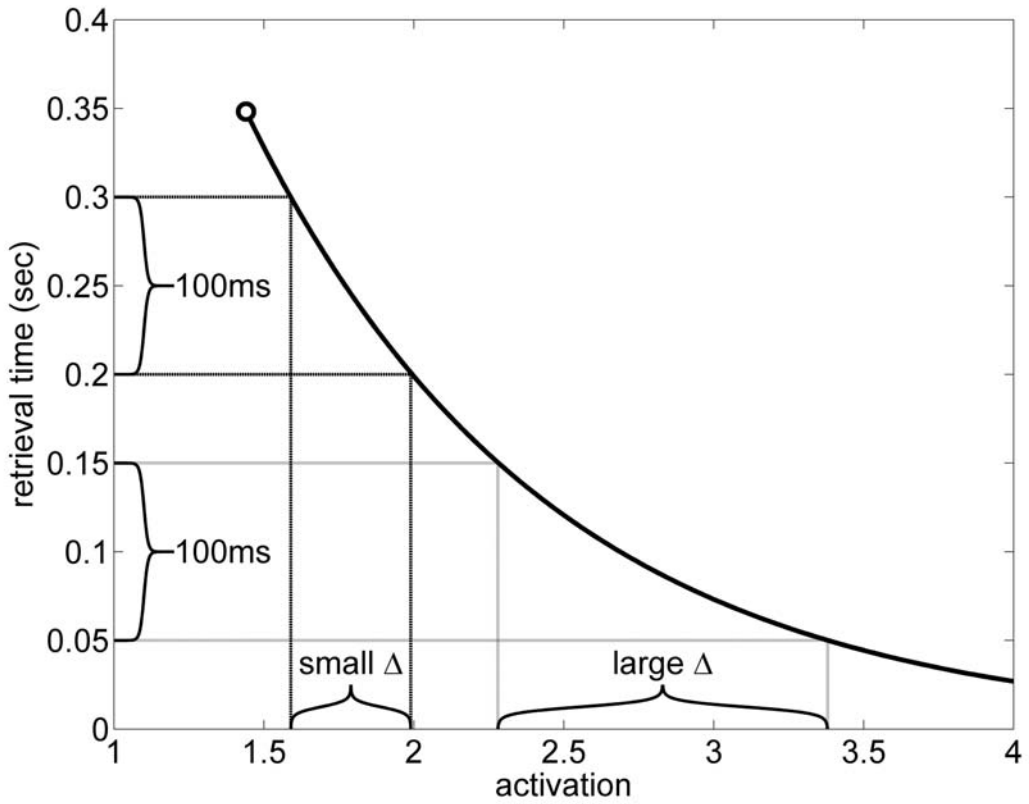

Figure 4. An exponential function relates a record's activation to its retrieval time. The open circle represents the retrieval time associated with τ, the retrieval criterion. Points to the left of τ are associated with low activation levels that result in retrieval failure.

We now turn to the implementation of the recognition and fluency heuristics within ACT-R, which depend on the probability and speed of retrieval, respectively.

The Recognition and Fluency Heuristics: Keys to Encrypted Frequency Information

In ACT-R, activation tracks environmental regularities, such as an object's frequency and recency of occurrence. Therefore, activation differences partly reflect frequency differences, which, in turn, may be correlated with a characteristic of the object, such as the population of a city. Thus, it may seem that inferences could be based on activation values read off the records. However, applications of ACT-R have long assumed that, just as the long-term potentiation of neurons in the hippocampus cannot be tapped directly, subsymbolic quantities such as activation cannot be accessed directly. We nevertheless propose that the system can capitalize on differences in activation associated with various objects by gauging how it responds to them. Two responses that are correlated with activation in ACT-R are (1) whether a record associated with a specific object can be retrieved and (2) how quickly the record can be retrieved. The first, *binary* response underlies our implementation of the recognition heuristic, and the second, *continuous* response underlies our implementation of the fluency heuristic. We show that the heuristics can be understood as tools that indirectly tap the environmental frequency information locked in the activation values. The heuristics' effectiveness depends on the strength of the chain of correlations linking the criteria, environmental frequencies, activations, and responses. As we will demonstrate, forgetting strengthens this chain. Before we describe the simulations, a more general remark about the notion of recognition in ACT-R is in order.

An ACT-R Model of Recognition

In modeling the recognition and fluency heuristics, we borrow from Anderson, Bothell, Lebierre and Matessa's (1998) account of recognition. In an episodic recognition task a person decides whether an item, typically a word has been encountered in some specific context, say, in a newspaper article. The responses of Anderson et al.'s (1998) model are determined by whether or not various memories are retrieved, thus the model assumes an all-or-none or high-threshold notion of recognition, which is consistent with how Goldstein and Gigerenzer's (2002) treat recognition.

In the literature on recognition memory, there is debate about whether such high-threshold models are compatible with the receiver operating characteristics (ROC) curves typically observed in recognition memory experiments (e.g., Batchelder, Riefer and Hu, 1994; Kinchla, 1994; Malmberg, 2002). ROC curves, which are diagnostic of a participant's ability to distinguish between different kinds of stimuli, can be derived by manipulating participants' response bias. Specifically, in a recognition memory experiment, they are encouraged to be more or less liberal in their tendency to say that they recognize a stimulus. Based on these judgments, one can plot for each level of response bias a point that corresponds to the resulting hit rate and false alarms rate. The problem with standard implementations of discrete-state models is that they yield linear ROC's curves. The curves participants generally produce, however, are curvilinear, and are consistent with the more widely accepted signal detection theory (SDT) view of recognition, in which memory judgments are based on continuous information.

Does this property of high-threshold models by extension disqualify the Anderson et al.'s (1998) recognition model? In fact, their model does not produce ROC curves of any sort, simply because no mechanisms were specified to handle changes in response bias or to generate confidence ratings (which can also be used to generate ROC curves). Although it is the case that straightforward modifications, such as varying the propensity of the model to

respond that it recognizes items, will not yield curvilinear ROC curves, one ought not jump to the conclusion that Anderson et al.'s model cannot produce appropriate ROC curves. Malmberg (2000) demonstrated that whether or not a high-threshold model produces linear or curvilinear ROC curves depends critically on the assumptions made about the mechanisms that produce the confidence ratings. So while ACT-R belongs to the class of high-threshold models, its retrieved memory records rest on a continuous memory variable (i.e., activation). This variable, in turn, could be used to construct (curvilinear) ROC curves based on confidence ratings.

To conclude, Anderson et al. (1998) did not consider basic receiver operating characteristics (ROC) curves. Yet, the model has accounted quite successfully for many recognition memory effects, including the vexing list-strength effect (Ratcliff, Clark, & Shiffrin, 1990), which could not be handled by the mathematical models existing at the time. In addition, the same basic mechanisms have been applied to dozens of empirical results in a wide range of domains. All in all, we believe we are on pretty solid theoretical ground by drawing on Anderson et al.'s ACT-R account of recognition for modeling the recognition and the fluency heuristic. In the discussion, we return to the distinction of binary versus continuous notions of recognition and consider a model that adopts a signal detection view of recognition.

Simulations of the Recognition and Fluency Heuristics

Figure 1 illustrates the basic steps in our simulations that applied our ACT-R models of the recognition and fluency heuristics to the city population comparison task. *First*, we constructed environments that consisted of the names of German cities and the days on which they were encountered. *Second*, the model learned about each city from the constructed environments by strengthening memory records associated with each city according to Equation 3, ACT-R's base-level activation equation. *Third*, we determined the model's recognition rates by fitting Equation 4, the probability of retrieving a record given its activation, to the rates at which Goldstein and Gigerenzer's (2002) participants recognized the cities. *Fourth*, these recognition rates were used to drive the performance of the recognition and fluency heuristics on the city population comparison task. We now describe these simulations in detail.

How The City Environments Were Constructued

In the simulations, the probability of encountering a German city name on a given day was proportional to its relative frequency in the *Chicago Tribune.* The frequencies were taken from Goldstein and Gigerenzer's (2002) counts of how often the 83 largest German cities were mentioned between January 1, 1985, and July 31, 1997. Thus, the probability of encountering city *i* on any given day was:

$$P(i) = \frac{f_i}{w}, \qquad (6)$$

where f_i is the total number of citations for the i^{th} city, and w is the total number days in the sample (the historical window). For example, Berlin, the largest city, was mentioned 3,484 times in the 4,747-day sample, so its daily encounter probability was .73. Duisburg, the 12th largest city, was mentioned 53 times, yielding a probability of .011. Based on these encounter probabilities, a historical environment was created for each city that consisted of a vector of 1s and 0s, where a 1 indicated that the city's name had been encountered on a particular day and 0 indicated that it had not. Because the size of the historical window is arbitrary, we set it to 4,747, the total number of days in Goldstein and Gigerenzer's analysis.

For this set of simulations, the probability of encountering a city on any given day was fixed according to Equation 6. That is, the probability of encountering a city was independent of when it was last encountered, though, of course, cities with higher probabilities would tend to have shorter lags between encounters than would less frequent cities. Later we report simulations that used environments with a more refined statistical structure, which led to comparable conclusions.

How The Activations for the Cities' Records Were Learned

As in the example illustrated in Figures 2 and 3, each city had an associated memory record that was strengthened according to Equation 3. When the end of each time window was reached, activation values for each city were calculated by averaging its activation across 500 constructed environments. The subsequent simulations are based on these average activation values. As one interprets the simulation results, however, it may be helpful to keep in mind Anderson's (1993) approximation to Equation 3:

$$B = k + \ln n - d \ln T \qquad (7)$$

where k is a constant, n is the number of times the item associated with the record has been encountered, and T is how long it has been since the record was first created. Taking the natural log compresses larger numbers more than smaller ones. Thus, because of this compression, each successive encounter with an item contributes less than the preceding encounter to the total activation. Similarly, activation decays quickly at first and more slowly thereafter, because each subsequent "tick" of the clock is compressed, and so subtracts less and less from the total activation.

Does Activation Capture the Correlation Between Environmental Frequency and the Criterion?

For the recognition and fluency heuristics to be useful inference strategies, activation needs to reflect the relation between objects' frequencies in the environment and their values on the criterion. Given that the ecological correlation between the raw citation counts and the city population size (r= .82) is high, judgments of city size based on citation counts can reasonably be taken as an upper bound on inferential performance. Indeed, inferences about which of two cities is larger based on these counts (where a higher count implies a larger city) are accurate in 76.5% of city comparisons. In comparison, inferences based on city's average activation (where a higher activation implies a larger city) have an accuracy of 76.4%, just .1 percentage point below those based on the raw frequency information. Thus, activation seems to closely track the cities' environmental frequencies. With these bounds on accuracy in mind, we now examine the performance of the recognition and fluency heuristics.

An ACT-R Model of the Recognition Heuristic

The performance of the recognition and fluency heuristics depends on the cities' recognition rates. In line with Anderson et al. (1998) a city was recognized when the record associated with the city could be retrieved. A city's recognition rate was estimated by fitting Equation 4, which relates a record's activation to its probability of retrieval, to the empirical recognition rates that Goldstein and Gigerenzer (2002) observed. Equation 4's two free parameters were estimated using the nonlinear regression function from SPSS 11.0. These are τ, the retrieval criterion (estimated to be 1.44 units of activation), and s, the activation noise (estimated to be .728). The correlation between the estimated and empirical recognition rates was high (r = .91). Figure 5 plots recognition as a function of activation. The points represent the empirical

recognition rates, and the S-shaped curve is the estimated recognition rates based on ACT-R's retrieval mechanisms (Equation 4).

To see how the Chicago students would be expected to do on the city comparison task, if they were to employ the recognition heuristic, we calculated the recognition heuristic's performance based on the empirical recognition rates recorded by Goldstein and Gigerenzer (2002). That is, if only one city was recognized, that city was chosen; otherwise, a guess was made. Performance of the recognition heuristic based on the empirical and model's recognition rates on all possible city pairs was .606 and .613 respectively, indicating good agreement between the behavior of the ACT-R model of the recognition heuristic and that expected from the students.

Based on this correspondence, we can now pose novel questions concerning whether the recognition and fluency heuristics benefit from loss of information in memory.

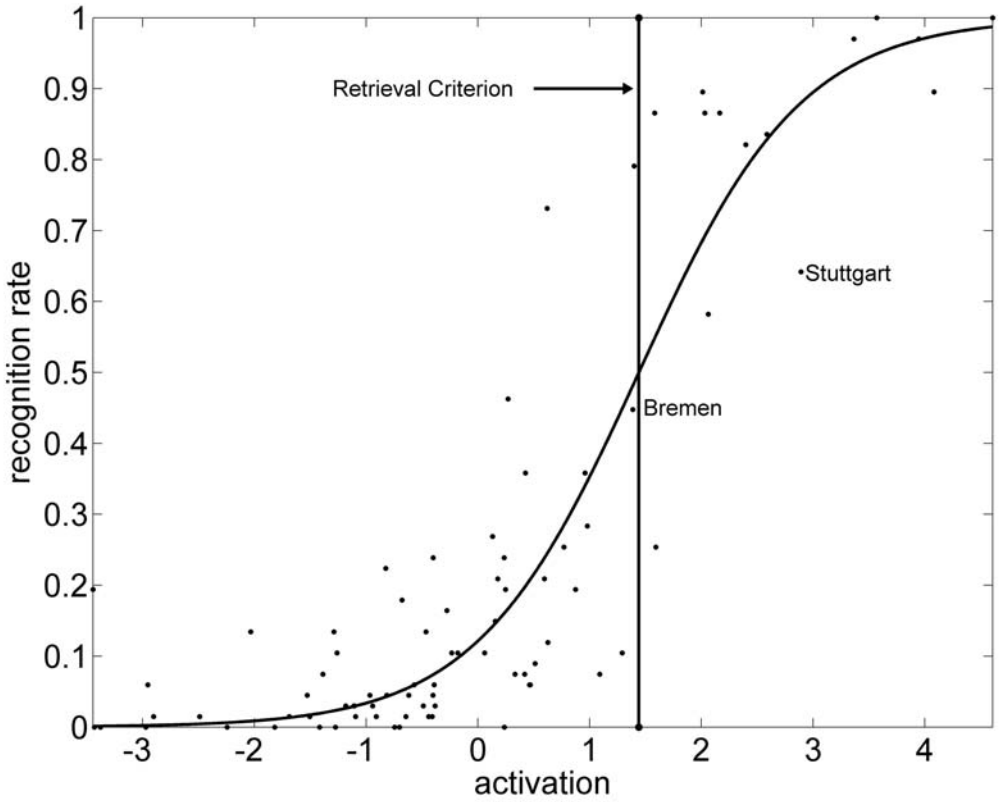

Figure 5. Recognition rate plotted as a function of activation. The points indicate the observed recognition rates of the 83 German cities. The S-shaped curve relates the activation of a city's record to its estimated recognition rate. For instance, Bremen has an observed recognition rate of .45, an activation of 1.39, and an estimated recognition rate of .48. Stuttgart has an observed recognition rate of .64, an activation of 2.89, and an estimated recognition rate of .88.

Does Forgetting Benefit the Recognition Heuristic?

To address this question, we varied the decay rate *d* (holding both the retrieval criterion, τ, and the activation noise, *s,* constant) and observed how the resulting changes in recognition affect inferences in the city population task.[5] The upper bound of the decay rate, 0, means no forgetting, the strength of a memory record is strictly a function of its frequency. Negative values of *d* imply forgetting, and more negative values imply more rapid forgetting. Using a

step size of .01, we tested *d* values ranging from 0 to –1, the latter being twice ACT-R's default decay rate. In Figure 6, the solid line shows the recognition heuristic's average performance on pair-wise comparisons of all German cities with more than 100,000 residents, including pairs in which it had to guess because both cities are recognized or unrecognized. Three aspects of this function are noteworthy. First, the recognition heuristic's performance assuming no forgetting (56% correct) is substantially worse than its performance assuming the "optimal" amount of forgetting (63.3% correct). Second, ACT-R's default decay value of –.5 yields 61.3% correct, only slightly below the peak performance level, which is reached at a decay rate of –.34. Third, the sensitivity curve has a flat maximum, with all decay values from –.13 to –.56 yielding performance in excess of 60% correct.

In other words, forgetting enhances the performance of the recognition heuristic, and the amount of forgetting can vary over a substantial range without compromising the heuristic's good performance. If there is too much forgetting (resulting in a situation in which most cities are unrecognized), however, the performance of the recognition heuristic eventually approaches chance level.

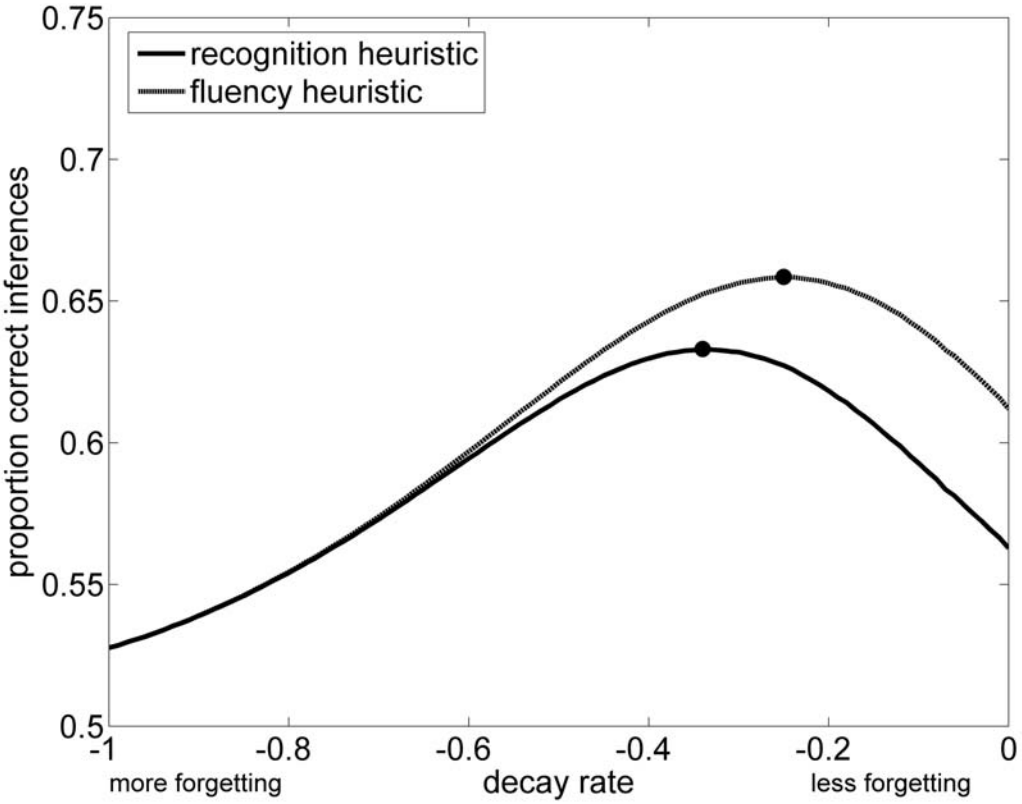

Figure 6. Proportion of correct inferences made by the recognition and fluency heuristics on all comparisons of the 83 largest cities in Germany. The amount of forgetting in the system was varied from 0, corresponding to no forgetting, and of –1, a high forgetting rate. The peaks of each curve are marked with dots.

How Does Forgetting Help the Recognition Heuristic's Performance?

Two quantities shed more light on the link between forgetting and the recognition heuristic. The first is the proportion of comparisons in which the recognition rule can be used as the basis for making a choice, that is, the proportion of comparisons in which only one of the cities is recognized. In Figure 7, the solid line shows that for the recognition rule this

application rate peaks when *d* equals –.28, an intermediate level of forgetting. The second quantity is the proportion of correct inferences made by the recognition heuristic in those choices to which it is applicable. As shown in Figure 8, this *recognition validity* generally increases with the amount of forgetting, peaking when *d* equals –1. The performance (Figure 6) and application rate (Figure 7) peak at nearly the same forgetting rates of -.34 and -.28, compared to the peak of -1 for the validity curve (Figure 8). So the decay rate of -.34 can be thought of as the optimal trade-off between the effects of forgetting on application rate and validity, with the application rate having the greater sway over performance. Thus, intermediate amounts of forgetting increase the performance of the recognition heuristic mostly by sharply increasing its applicability and, to a lesser extent, by increasing its validity.

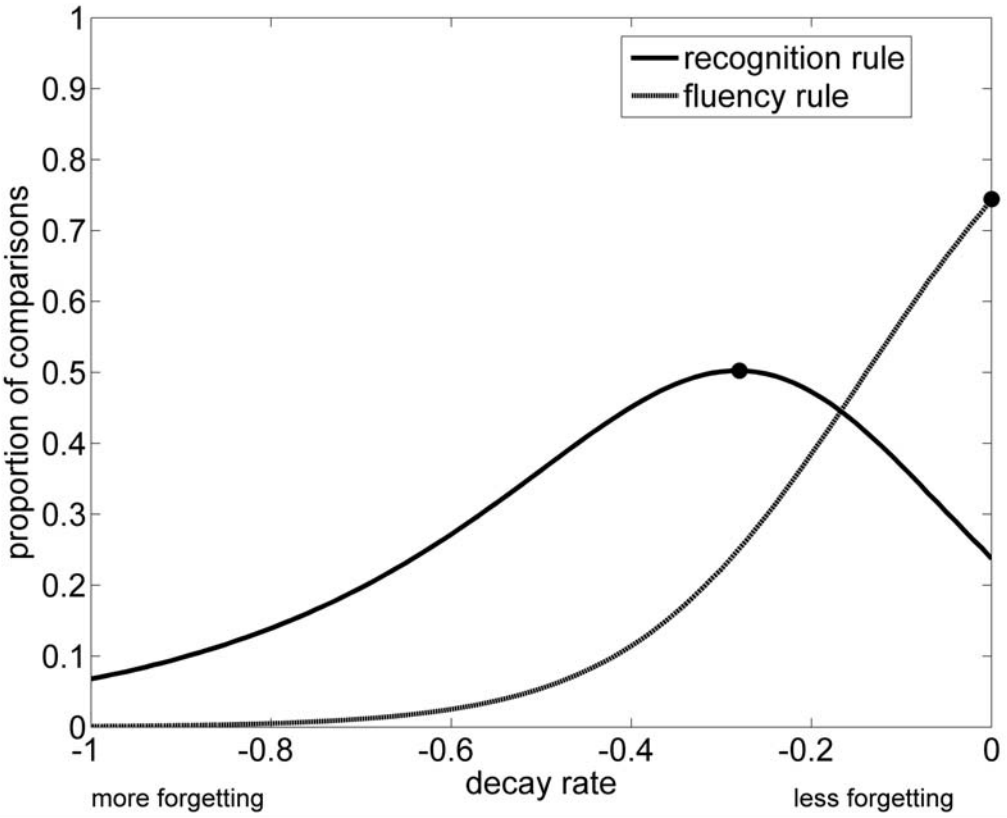

Figure 7. Application rate of the recognition rule and fluency rule. The application rate for the recognition rule is the proportion of all comparisons in which only one of the cities is recognized. The application rate for the fluency rule is the proportion of all comparisons in which both cities are recognized.

The results of the ACT-R simulations of the recognition heuristic suggest that forgetting serves to maintain the memory system's partial ignorance, a precondition for the heuristic's functioning. Loss of some information—a loss that is not random but a function of a record's environmental history— fosters the performance of the recognition heuristic. But how robust is this result and is it limited to the recognition heuristic that takes recognition to be all-or-one? To find out whether the phenomenon generalizes to memory-based inference strategies that make finer distinctions than that between recognition and nonrecognition, we now turn to the fluency heuristic.

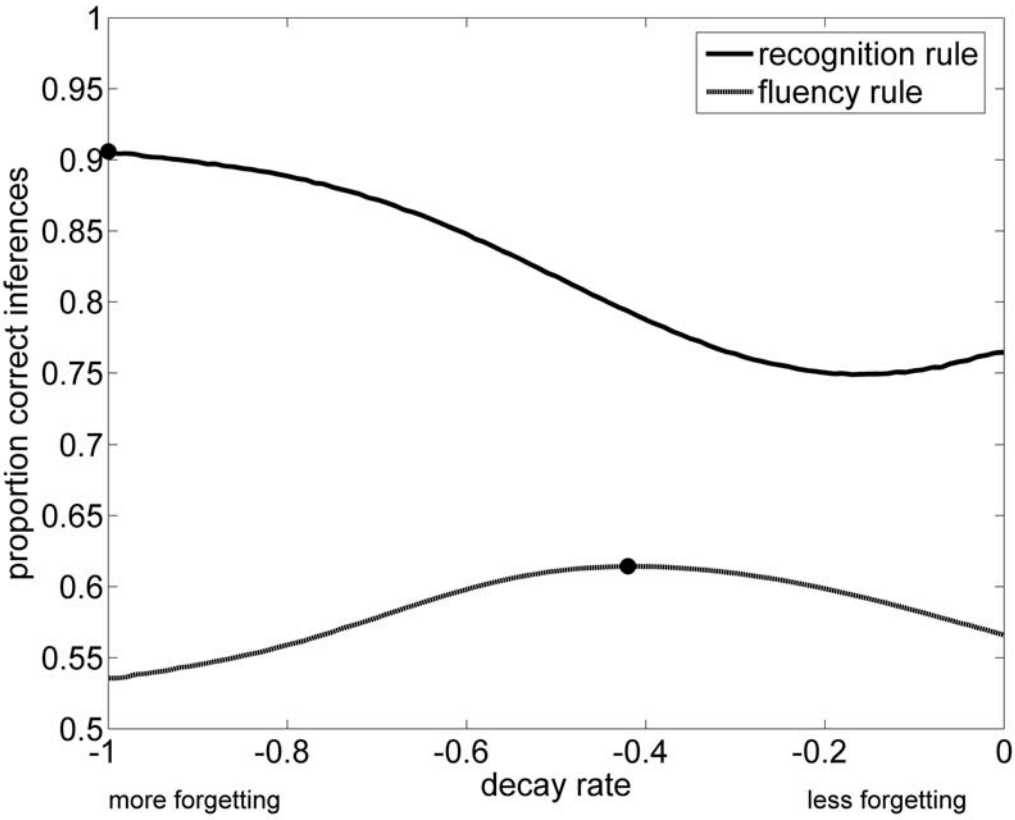

Figure 8. Validity of the recognition rule and fluency rule. The validity of a rule is the proportion of correct inferences that the rule makes when it can be applied.

An ACT-R Model of the Fluency Heuristic

The recognition heuristic depends on the correlation between recognizing an object and its environmental frequency, but when both objects are recognized this correlation is of no use. Yet, recognized objects could differ in their activation levels, indicating a difference in frequency. Although these activation differences cannot be assessed directly, Equation 5 raises the possibility that retrieval time, because of its one-to-one mapping with activation, could be used as a proxy for activation. To see how this might be accomplished, let us assume that people are sensitive to differences in recognition times (i.e., retrieval times). Specifically, let us suppose that the American students in Goldstein and Gigerenzer's (2002) studies could tell the difference between instantaneously recognizing Berlin, for instance, and taking a moment to recognize Stuttgart. We suggest that such differences in recognition time partly reflect retrieval time differences, which, according to Equation 5, reflect the base-level activations of the corresponding memory records.

Moreover, retrieval time allows us to make the notion of fluency of reprocessing more precise. Operationalizing fluency as retrieval time, we now implement the fluency heuristic within the ACT-R framework. In Table 2, we specify the fluency heuristic for the two-alternative choice between two cities as a set of three production rules that build on the rules constituting the recognition heuristic presented in Table 1.

Table 2. The production set that implements the Fluency Heuristic

Guessing rule
If the goal is to choose the larger of two cities, City X and City Y,
And a record with the name City X ***cannot*** be retrieved
And a record with the name City Y ***cannot*** be retrieved
Then set a goal to guess

Recognition rule
If the goal is to choose the larger of two cities, City X and City Y,
And a record with the name City X ***can*** be retrieved
And a record with the name City Y ***cannot*** be retrieved
Then set a goal to respond that City X is the larger city

Fluency rule
If the goal is to choose the larger of two cities, City X and City Y,
And a record with the name City X ***can*** be retrieved
And a record with the name City Y ***can*** be retrieved
Then set a goal to compare retrieval times and respond that the city retrieved fastest is larger

The first two rules embody the essential components of the recognition heuristic: Guess when neither alternative is recognized, and choose the recognized alternative when one is recognized and the other is not. Triggered only when both alternatives are recognized, the *fluency rule* sets a goal of comparing retrieval times. Hereafter we refer to the complete set of rules in Table 2 as the fluency heuristic and to the third rule of the set specifically as the *fluency rule*.

In the interest of psychological plausibility, we built in limits on the system's ability to discriminate between retrieval times. Rather than assuming that the system can discriminate between minute differences in any two retrieval times, we assume that if the retrieval times of the two alternatives are less than a just noticeable difference (JND) apart, then the system must guess. Guided by Fraisse's (1984) conclusion on the basis of an extensive review of the timing literature that durations of less than 100 ms are perceived as instantaneous, we set the JND to 100 ms rather than modeling the comparison of retrieval times in detail. We do not claim, however, that this value captures people's actual thresholds exactly.

Comparison of the Fluency and Recognition Heuristics

The fluency heuristic assumes that people compare the retrieval times for the two objects and choose the object that is more quickly recognized (the other object may be more slowly recognized or come up unrecognized). How accurate is this heuristic compared with the recognition heuristic? Surprisingly, using the default decay rate of –.5, the fluency heuristic (62.1%) performs only slightly better than the recognition heuristic (61.3%).

Let us analyze this performance in more detail. Recall that recognition validity is the probability of getting a correct answer when one object is recognized and the other is not. The recognition validity in our simulation was .82. The overall accuracy of the recognition heuristic is reduced, because the heuristic resorts to guessing in cases in which both cities are recognized (5.5% of all comparisons) or both cities are not recognized (58.2% of all

comparisons). Analogous to the recognition validity, *fluency validity* (i.e., the validity of the fluency rule) is the probability of getting a correct answer when both objects are recognized. The fluency validity is .61, lower than the recognition validity but still higher than the recognition heuristic's chance-level performance when recognition does not discriminate between objects.

From this it follows that the fluency heuristic's competitive advantage over the recognition heuristic depends on the relative frequency of city pairs in which both objects are recognized. It is only in these comparisons that the performance of the two heuristics can differ; when one object is recognized and the other is not, the two heuristics behave identically, and when neither object is recognized, both must guess. This conclusion is illustrated by comparing the performance of the two heuristics on different subsamples of cities. For example, if comparisons are restricted to the 10 largest cities (resulting mostly in pairs in which both objects are recognized), the fluency heuristic has about a 5% performance advantage over the recognition heuristic (63.8% vs. 58.8%). This advantage drops to 3% when the 20 largest cities are included (66.7% vs. 63.3%) and to less than 1% when all 83 of the largest cities are included. In short, the fluency heuristic compares most favorably with the recognition heuristic when the sample is dominated by large cities that tend to be easily recognized.

Does Forgetting Benefit the Fluency Heuristic?

Loss of information bolsters the performance of the recognition heuristic, but does it give a boost to the fluency heuristic as well? Indeed, the dashed line in Figures 6 shows that performance of the fluency heuristic peaks at a decay rate of –.25. How can it be that the fluency heuristic's performance peaks at intermediate levels of forgetting—a heuristic that feeds on recognition knowledge and not lack thereof (as the recognition heuristic does)? Is it possible that the peak in performance at intermediate levels of forgetting stems solely from the recognition rule within the fluency heuristic?

To investigate whether the fluency rule enjoys any independent benefit of forgetting, we analyzed the set of city comparisons in which both cities are recognized (to which the fluency rule applies) and the proportion of correct inferences that the fluency rule makes in this set as a function of forgetting. Figures 7 shows that the fluency rule's application rate drops as forgetting rises, as one would expect given that the fluency rule applies only when both cities are recognized. When it applies, however, the fluency rule indeed benefits from intermediate levels of forgetting. As Figures 8 demonstrates, the fluency rule's validity peaks at the intermediate decay rate of -.42, though this peak is well below that of the recognition rule's validity. That is, the peak in the fluency heuristic's performance at intermediate levels of forgetting stems from benefits of forgetting that cannot be reduced to those for the recognition rule. But how does the fluency rule benefit from forgetting?

What Causes the Fluency Rule's Validity to Peak at Intermediate Decay Rates?

To understand how one important factor contributes to the shape of the fluency rule's validity curve, let us revisit the exponential function that relates activation to latency in Figure 4. Consider first retrieval times of 200 and 300 ms, which correspond to activations of 1.99 and 1.59 respectively. For these relatively low activations, only a small difference of .40 units of activation is required to exceed the 100 ms JND. In contrast, the 100 ms difference in retrieval time between 50 and 150 ms corresponds to a difference of 1.1 units of activation. Thus, by shifting the activation range downward, forgetting helps the system settle on activation levels corresponding to retrieval times that can be more easily discriminated. In the

case of the fluency heuristic, memory decay prevents the activation of (retrievable) records from becoming saturated.

Less is More Even for the Fluency Heuristic

Intermediate amounts of forgetting benefit not only the recognition heuristic but the fluency heuristic as well. Generally, the application and validity rates of the recognition rule and the validity rate of the fluency rule profit from faster forgetting, while the application rate of the fluency rule tugs strongly toward slower forgetting. In short, three of the four quantities that determine the performance of the fluency heuristic peak at faster decay rates. This observation is akin to the less-is-more effect for the recognition heuristic (see Figures 2 and 3 in Goldstein and Gigerenzer, 2002). In the less-is-more context, decision makers who know less can exhibit greater inferential accuracy than do those who know more. In the present context, decision makers who have intermediate rates of memory decay can make more accurate inferences than those with little or no decay—whether they are using the recognition heuristic or the fluency heuristic.

Is the Beneficial Effect of Forgetting Robust in Environments with Natural Clustering?

In the simulations reported thus far, the probability of a city's name being mentioned on a particular day was taken to be proportional to its overall citation rate and independent of when it was last mentioned. This assumption ignores a potentially important aspect of the environmental structure, namely, that the occasions on which an item (e.g., a city name) is encountered tend to cluster temporally. Consider, for instance, the pattern for Chicago in Figure 2. The cluster in early 1987 relates to stories about the mayoral election. The cluster toward the end of 1987 relates to the mayor dying in office. Even without knowing what a word means, one can predict well how likely it is to be mentioned in the future based on how recently and frequently it has been mentioned in the past. ACT-R's activation equation was designed to be sensitive to just such patterns (Anderson & Schooler, 1991). To find out whether the benefits of forgetting generalize to a model of the environment that reflects the natural clustering of events, the models already presented were run again, but this time on environments that contained natural clustering. This was achieved by making the probability of encountering a city name dependent both on how long ago it was last encountered and how frequently it was encountered in the recent past. These dependencies were modeled on those found by Anderson and Schooler (1991) in their analysis of word usage in the *New York Times* headlines.

Figure 9 shows how the performance of the recognition and the fluency heuristics depends on the decay rate in this clustered environment. Although these curves are rough-hewn because they are based on 11 decay values as opposed to the 101 decay values used to map out the other decay functions, the results are consistent with those of Figure 6, in which the probability of a word's being mentioned on a given day was proportional to its overall environmental frequency (see Equation 5). The similarity between the two sets of results suggests that learning over extended periods smoothes out the possible effects of clustering on performance.

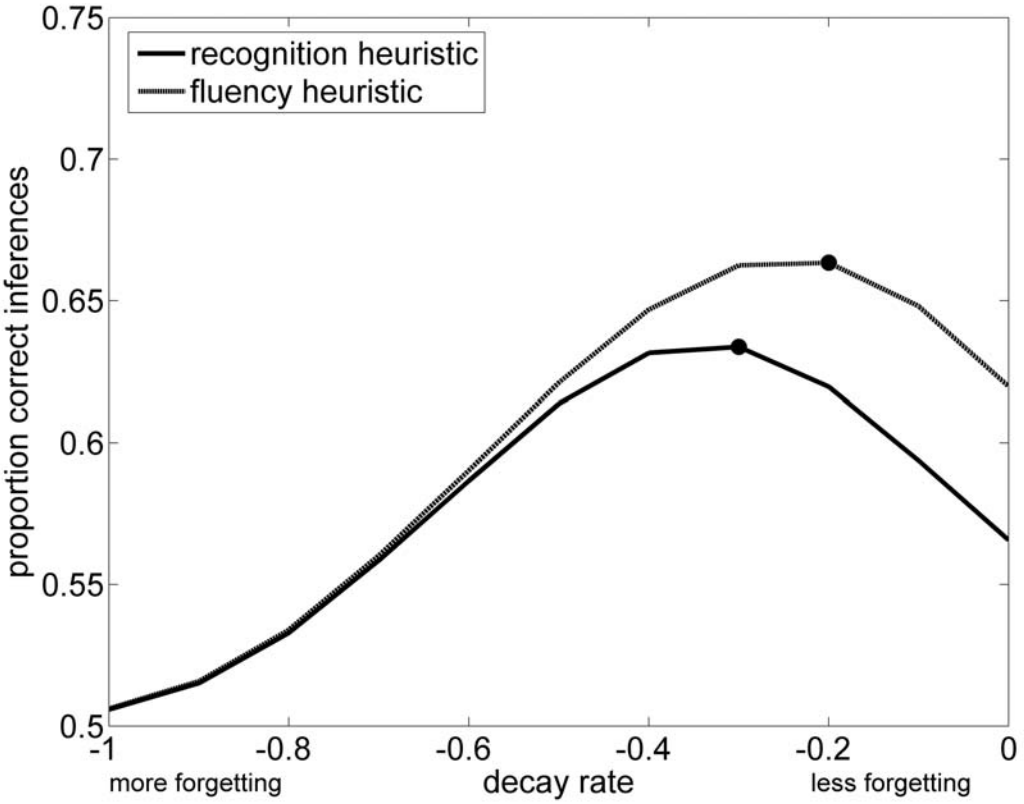

Figure 9. Performance of the recognition and fluency heuristics as the amount of forgetting varies, when activations are based on simulated environments that display more natural clustering of city mentions.

General Discussion

Some theorists have argued that forgetting is indispensable to the proper working of memory. Building on the notion of beneficial forgetting, we have demonstrated that ecologically smart loss of information---loss that is not random but reflects the environmental history of the memory record---may not only foster memory retrieval processes but can also boost the performance of inferential heuristics that exploit mnemonic information such as recognition and retrieval fluency. We did so by implementing inferential heuristics within an existing cognitive architecture, thus enabling us to analyze how parameters of memory such as information decay affects inferential accuracy. This analysis also revealed three distinct reasons for why forgetting and heuristics can work in tandem. In the case of the recognition heuristic, intermediate amounts of forgetting maintain the systematic partial ignorance on which the heuristic relies and increase somewhat the heuristic's validity, the probability that it correctly picks the larger city. In the case of the fluency heuristic, intermediate amounts of forgetting boost the heuristic's performance by maintaining activation levels corresponding to retrieval latencies that can be more easily discriminated. In what follows, we (1) discuss the robustness of the beneficial effects of forgetting, (2) investigate how the fluency heuristic relates to the availability heuristic, (3) discuss whether it is worthwhile to maintain the distinction between the fluency and recognition heuristics, and (4) conclude by examining whether forgetting plausibly could have evolved to serve heuristic inference.

A Signal-Detection View of Recognition: How Robust are the Beneficial Effects of Forgetting?

We see Goldstein and Gigerenzer's (2002) recognition heuristic not so much as a model of recognition, but rather as a model of how the products of the recognition process could be used to make decisions. Given that the recognition heuristic starts where models of recognition leave off, it seems reasonable, for this purpose, to assume that items are either recognized or they are not, even if this is a simplification. However, as our interest was in the impact of forgetting on how the recognition heuristic operates, we needed to consider the details of the underlying recognition process. We adapted Anderson, Bothell, Lebiere, and Matessa's (1998) ACT-R model of episodic recognition, which is a high-threshold model that depends on the all-or-none retrieval of appropriate memory chunks. By adding fluency our model is no longer strictly a high-threshold model, because there is continuous information available for the retrieved items, though not for those items that failed to be retrieved. Models like this were considered by Swets, Tanner, and Birdsall (1961).

In what follows, we investigate whether our results are robust in the context of a signal-detection view of recognition memory, currently the most widely shared view of recognition. In this view, there is a potential for discriminability even among unrecognized items. Signal detection theory describes a decision maker who must choose between two (or more) alternatives—for instance, whether or not she has encountered a present stimulus previously—on the basis of ambiguous evidence (Green & Swets, 1966). This uncertain evidence is summarized by a random variable that has a different distribution under each of the alternatives (encountered versus not encountered). The evidence distributions overlap, thus some events are consistent with each of the two alternatives. The decision maker establishes a decision criterion *C* that divides the continuous strength of evidence axis into regions associated with each alternative, for instance, the "recognized" versus the "unrecognized" region. If the evidence value associated with an event in question exceeds *C*, the decision maker will respond "recognized"; otherwise, she will respond "unrecognized." On this view, though people's decisions are dichotomous (recognized versus unrecognized), the underlying recognition memory and strength-of-evidence axis are not. Moreover, the unrecognized items are not of one kind but differ in gradation of strength, thus affording discrimination even if items are not recognized.

In the present ACT-R models of the fluency and recognition heuristic, the retrieval criterion, τ, doubles as a decision criterion for recognition. This dual role for τ is consistent with how it was used by Anderson, Bothell, Lebiere, & Matessa (1998). By decoupling τ's functions, we can now implement a version of the fluency heuristic that attempts to distinguish between unrecognized items. Specifically, if the retrieval criterion is assumed to be *lower* than the recognition decision criterion, then the fluency rule will apply to comparisons in which both objects exceed the modest retrieval criterion but remain unrecognized. The fluency heuristic can then capitalize on the fact that one unrecognized name is perhaps more fluently processed (i.e., has a higher activation value and faster retrieval time within ACT-R) than the other unrecognized name.

Will the benefits of forgetting generalize to this version of the fluency heuristic? To answer this question, we reran the simulations of the fluency heuristic but set τ so low that all memory records would be retrieved. As a result, all decisions were handled by the fluency rule. Figure 10 shows that forgetting also facilitates the performance of this version of the fluency heuristic. As in the previous simulations of the fluency heuristic, the reason for the performance boost is that loss of information lowers the range of activation to levels

corresponding to more discriminable retrieval times. In other words, a given difference in activation in a lower part of the range results in a larger, more detectable difference in retrieval times than does the same-sized difference in a higher part of the range. Thus, the beneficial effects of forgetting also proves robust in a signal detection view of recognition memory.

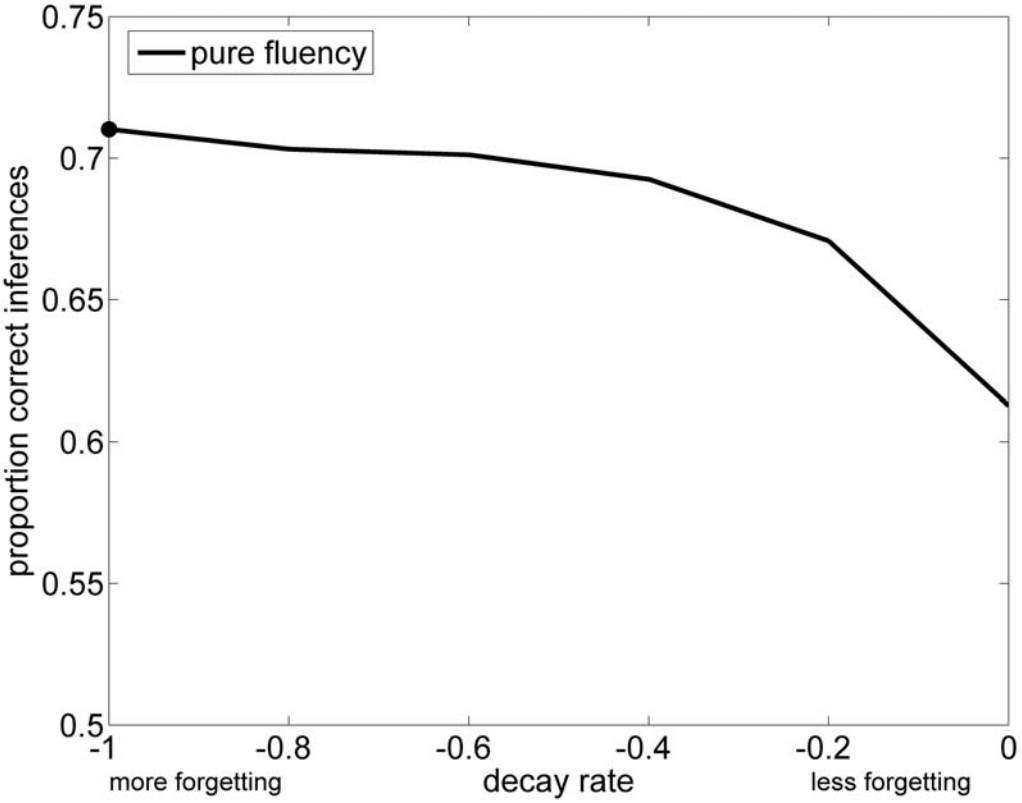

Figure 10. Performance of the fluency heuristic when continuous information is available for all city records.

The Fluency and Availability Heuristics: Old Wine in a New Bottle?

The fluency heuristic feeds on environmental frequencies of occurrences that are related to criterion variables such as population size. It thus can be seen as another ecologically rational cognitive strategy belonging to the *adaptive toolbox* of fast and frugal heuristics (Gigerenzer et. al, 1999). But is it new? Fluency shares an important property with one of the three major heuristics investigated in the heuristics-and-biases research program, namely, availability (Kahneman, Slovic, & Tversky, 1982). Both the availability heuristic and the fluency heuristic capitalize on a subjective sense of memory fluency. Tversky and Kahneman (1973) suggested that people using the availability heuristic assess the probability and the frequency of events on the basis of the *ease* or the *frequency* with which relevant instances of those events can be retrieved from memory. Thus, they proposed two notions of availability (Tversky & Kahneman, 1973, pp. 208, 210), one that depends on the actual frequencies of instances retrieved and one that depends on the ease with which the operation of retrieval can be performed (for more on the distinction between these two notions of availability, see Sedlmeier, Hertwig, & Gigerenzer, 1998; Hertwig, Pachur, & Kurzenhäuser, 2004).

If one understands availability to mean ease of retrieval, then the question arises how ease should be measured. Sedlmeier et al. (1998), for example, proposed measuring ease in terms of speed of retrieval. Thus interpreted, availability becomes nearly interchangeable with fluency, although the fluency heuristic retrieves the event itself (e.g., the name of a city) whereas the availability heuristic retrieves instances from the class of events (e.g., people who died of a heart attack versus people who died of lung cancer to estimate which of the two diseases has a higher mortality rate). We have no objection to the idea that the fluency heuristic falls under the broad rubric of availability. In fact, we believe that our implementation of the fluency heuristic offers a definition of availability that interprets the heuristic as an ecologically rational strategy by rooting fluency in the informational structure of the environment. This precise formulation transcends the criticism that availability has been only vaguely sketched (e.g., Fiedler, 1983; Lopes & Oden, 1991; Gigerenzer, 1996). In the end, how one labels the heuristic that we have called fluency is immaterial, because as Hintzman (1990) observed, "the explanatory burden is carried by the nature of the proposed mechanisms and their interactions, not by what they are called" (p. 121).

The Fluency and Recognition Heuristics: Are they the same thing?

Heuristics, at their core, are models of cognitive processes. So, asking whether the fluency and recognition heuristics are the same thing amounts to asking whether they process information identically. In terms of our ACT-R implementation the answer to this question is no. The recognition rule, once its conditions are matched, can proceed immediately to make a decision. The fluency rule, in contrast, entails the additional steps required to compare the retrieval times for the respective objects. Having two distinct rules improves the overall efficiency of the system, because information is processed only as much as is necessary to make a decision. But there is a second reason, independent of our implementation for keeping the heuristics separate. By assuming two heuristics, we can investigate situations where one heuristic may be more applicable and effective than the other. For instance, the recognition heuristic may be more robust in the face of time pressure. When there is not enough time for distinctions in degrees of recognition, or for comparisons thereof, coarser information such as whether items are recognized or not may do the job. But even when people do have enough time to evaluate familiarity (or fluency), there may be factors that affect their sense of fluency but are less strongly related, unrelated or even negatively related to recognition. For example, priming may be more likely to disrupt fluency assessments than recognition judgments. Moreover, if people had insight into the relative accuracy of recognition and fluency in a particular context, they may be able to select one heuristic over the other. By assuming two rather than one heuristic we retain the degrees of freedom to identify and model such situations.

What Came First: The Forgetting or the Heuristics?

One interpretation of the beneficial effect of forgetting as identified here is that the memory system loses information at the rate that it does in order to boost the performance of the recognition and fluency heuristics and perhaps other heuristics. On this view, an optimal amount of forgetting has evolved in the cognitive architecture in the service of memory-based inference heuristics. Though such a causal link may be possible in theory, we doubt that evolving inferential heuristics gave rise to a degree of forgetting that optimized their performance. The reason is that memory has evolved in the service of multiple goals. It is therefore problematic to argue that specific properties of human memory—for instance, forgetting and limited short-term memory capacity—have optimally evolved in the service of a single function. While such arguments are seductive—for an example, see Kareev's (2000)

conjecture that limits on working memory capacity have evolved "so as to protect organisms from missing strong correlations and to help them handle the daunting tasks of induction" (p. 401)—they often lack a rationale for assuming that the function in question has priority over others.

On what grounds can one say that, for example, induction, object recognition, correlation detection, classification, or heuristic inference is the most important cognitive function? In the absence of an analysis that supports a principled ranking of these functions or a convincing argument as to why forgetting would have evolved in the service of one single function (and then later may have been co-opted by others), we hesitate to argue that memory loses information at the rate that it does *in order to* boost the performance of heuristics. We find it more plausible that the recognition heuristic, the fluency heuristic, and perhaps other heuristics have arisen over phylogenetic or ontogenetic time to exploit the existing forgetting dynamics of memory. If this were true, a different set of properties of memory (e.g., faster or slower forgetting rate) could have given rise to a different suite of heuristics.

Future Steps

By linking two research programs—the program on fast and frugal heuristics and the ACT-R research program—we were able to ground inference heuristics that exploit mnemonic information in a cognitive architecture. We believe that this synthesis opens potential avenues of research that go beyond those reported here. By implementing other heuristics, one could, for instance, investigate to what extent the benefits of forgetting may generalize to other heuristics, such as Take The Best, that rely on complexes of declarative knowledge (e.g., "Munich has a professional soccer team"). In addition, such implementations may point toward other heuristics that have yet to be discovered.

We also believe that implementing heuristics within a cognitive architecture facilitates the investigation of questions that have been notoriously difficult to tackle within research on heuristics—issues such as how different heuristics are selected and how they are aquired. For example, Rieskamp and Otto (2003) have shown that associative learning mechanisms can capture how participants select between heuristics in ways that are adaptive for particular environments. Nellen (2003) found that associative learning mechanisms used in ACT-R can achieve an adaptive match between heuristics and environment structure as well. These investigations, however, presupposed the existence of a set of heuristics to select from. Stepping back even further, one may ask what are the "building blocks" of the fast and frugal heuristics, and what are the rules, the constraints, that govern the composition of the buidling blocks into new heuristics. Little progress, if any, has been made on this issue of the acquistion of heuristics. We believe that one promising place to look for building blocks and constraints on their composition is in the basic mechanisms of ACT-R. Building heuristics on an ACT-R foundation ensures, at the very least, that they are cognitively plausible. In addition, constructing heuristics in this way will enrich our understanding of the relation between the heuristics in the *adaptive toolbox* (see Gigerenzer et al., 1999) and their basic cognitive foundations.

Conclusion

Analyses of cognitive limits, a well-studied topic in psychology, are usually underpinned by the assumption that cognitive limits, such as forgetting, pose a serious liability (see Hertwig & Todd, 2003). In contrast, we demonstrated that forgetting might facilitate human inference performance by strengthening the chain of correlations, linking the criteria, environmental frequencies, activations and the speed and accuracy of fundamental memory retrieval

processes. The recognition and fluency heuristics, we argued, use the response characteristics of these basic memory processes as a means to indirectly tap the environmental frequency information locked in the activations. So in light of the growing collection of beneficial effects ascribed to cognitive limits (see Hertwig & Todd, 2003), we believe it timely to reconsider their often exclusively negative status, and to investigate which limits may have evolved to foster which cognitive processes, and, vice versa, which processes may have evolved to exploit specific limits—as we propose in the case of heuristic inference and forgetting.

References

Altmann, E. M. & Gray, W. D. (2002). Forgetting to remember: The functional relationship of decay and interference. *Psychological Science, 13,* 27-33.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science,* 2, 396-408.

Anderson, J. R., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review, 104,* 728–748.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96,* 703–719.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General, 128,* 186–197.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language, 38,* 341–380.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1120–1136.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Anderson, J. R., & Schooler, L. J. (2000). Adaptive memory. F. Craik and E. Tulving, (Eds.), *Oxford Handbook of Memory.* Oxford, UK: Oxford University Press.

Batchelder, W. H., Riefer, D. M., & Hu, X. (1994). Measuring memory factors in source monitoring: Reply to Kinchla *Psychological Review*, 101, 172-176.

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General, 121,* 446–458.

Bjork, E. L., & Bjork, R. A. (1988) On the adaptive aspects of retrieval failure in autobiographical memory. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.). *Practical Aspects of Memory II* (insert page numbers). London: Wiley.

Ebbinghaus, H. (1885/1964). *Memory: A contribution to experimental psychology.* Mineola, NY: Dover.

Estes, WK (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, 62, 145-154.

Fiedler, K. (1983). On the testability of the availability heuristic. In R. W. Scholz (Ed.), Decision making under uncertainty (pp. 109–119). Amsterdam: North-Holland.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669.

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

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

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, P. M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (pp. 37–58). New York: Oxford University Press.

Green, D. M., & & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hertwig, R., Gigerenzer, G., & Hoffrage, U. (1997). The reiteration effect in hindsight bias. *Psychological Review, 104,* 194–202.

Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2004). *Judgments of risk frequencies: Tests of possible cognitive mechanisms*. Manuscript submitted for publication.

Hertwig, R., & Todd, P. M. (2003). More is not always better: The benefits of cognitive limits. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment and decision making* (pp. 213–231). Chichester, England: Wiley.

Hintzman, D. L. (1990). Human learning and memory: Connections and dissociations. *Annual Review of Psychology, 41,* 109–139.

Jacoby L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception and concept learning. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 18, pp. 1–47). New York: Academic Press.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110,* 306–340.

Jacoby, L. L., Kelley, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality & Social Psychology, 56,* 326-338.

James, W. (1890). *The principles of psychology* (Vol. 1). New York: Holt.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgement under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review, 107,* 397–402.

Kelley, C. M., & Jacoby, L. L. (1998). Subjective reports and process dissociation: Fluency, knowing, and feeling. *Acta Psychologica, 98,* 127–140.

Kinchla, R.A. (1994). Comments on Batchelder and Riefer's multinomial model for source monitoring. *Psychological Review*, 101, 166–171.

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology, 51,* 481–537.

Lopes, L. L., & Oden, G. D. (1991). The rationality of intelligence. In E. Eels & T. Maruszewski (Eds.), *Poznan studies in the philosophy of the sciences and the humanities* (Vol. 21, pp. 225–249). Amsterdam: Rodopi.

Luria, A. R. (1968). *The mind of mnemonist*. New York: Basic Books.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(2), 380-387.

McGeoch, JA (1932). Forgetting and the law of disuse. *Psychological Review*, 39, 352-370

Nellen, S. (2003). The use of the "Take-the-Best" heuristic under different conditions, modeled with ACT-R. In: F. Detje, D Dörner and H. Schaub (eds.), *Proceedings of the fifth international conference on cognitive modeling* (pp. 171-176). Bamberg: Universitätsverlag Bamberg

Peterson, S., & Simon, T. J. (2000). Computational evidence for the subitizing phenomenon as an emergent property of the human cognitive architecture. *Cognitive Science, 24,* 93–122.

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science*, 27, 431-452.

Rieskamp, J., & Otto, P. E. (2003). SSL: A theory of how people learn to select strategies. Submitted for publication.

Schooler, L. J., & Anderson, J. R. (1997). The role of process in the Rational Analysis of Memory, *Cognitive Psychology*, 32, 219-250.

Schunn, C., & Anderson, J. R. (1998). Scientific discovery. In J. R. Anderson & C. Lebiere (Eds.). *The atomic components of thought* (pp. 255–296). Mahwah, NJ: Erlbaum.

Sedlmeier, P., Hertwig, R., & Gigerenzer, G. (1998). Are judgments of the positional frequencies of letters systematically biased due to availability? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 754–770.

Schwarz, N., & Vaughn, L. A. (in press). The availability heuristic revisited: Recalled content and ease of recall as information. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *The psychology of intuitive judgment: Heuristics and biases*. Cambridge, England: Cambridge University Press.

Todd, P. M., & Kirby, S. (2001). I Like What I Know: How Recognition-Based Decisions Can Structure the Environment. In J. Kelemen & P. SosÃk (Eds.), *Advances in Artificial Life : 6th European Conference*. Heidelberg: Springer-Verlag.

Toth, J. P., & Daniels, K. A. (2002). Effects of prior experience on judgments of normative word frequency: Automatic bias and correction. *Journal of Memory and Language, 46,* 845–874.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232.

Whittlesea, B. W. A. (1993). Illusions of familarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1235–1253.

Authors' Notes

Our thanks go to John Anderson, Gerd Gigerenzer, Kenneth Malmberg Carmi Schooler, Peter Todd , John Wixted and the members of the ABC Research Group for their many constructive comments. We also thank Dan Goldstein for providing us with recognition rates and newspaper citation counts for the German cities. Please send correspondence to Lael Schooler, Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, Lentzeallee 94, 14195 Berlin, Germany. Electronic mail may be sent to schooler@mpib-berlin.mpg.de.

Footnotes

[1] This score of correct inferences is derived arise as follows: There are a total of 190 comparisons. In 45 of them, the middle brother resorts to guessing because he recognizes none, thus scoring 50% correct inferences. In another 45, in which he recognizes both, he scores 60% correct (the validity of his additional knowledge). In the remaining 100 comparisons, he scores 80% correct inferences (due to the validity of his recognition knowledge). Thus, the total score of correct inferences equals 45*.5 + 45*.6 + 100*.8 = 129.5/190 = 68%.

[2] In previous publications, Anderson and colleagues have called this need probability.

[3] Within ACT-R time dependent forgetting is attributed to memory *decay*. In contrast, many memory researchers, beginning with McGeoch (1932), have argued that tying forgetting to the passage of time through decay is simply a re-description of the empirical phenomenon, rather than a description of an underlying process. As an alternative to decay, researchers have typically favored explanations that attribute time dependent forgetting to *interference* (e.g., Estes, 1955; Raaijmakers, 2003). On the interference view of forgetting, memory for a stimulus gets encoded in a particular context, consisting of myriads of internal and external elements. Those, in turn, have the potential to later act as retrieval cues. The context, however, "drifts" as cues (randomly) enter and leave it. As a consequence, fewer cues are shared between the original encoding and the retrieval contexts over time, thus lowering the probability of recall of the target stimulus. Within the ACT-R framework memory decay is claimed to be functional, but at the same time there is no commitment to the underlying causes of memory decay. Thus, ACT-R does not preclude explaining decay as the aggregate result of factors such as contextual drift process, neural degradation, or some other causes altogether. In addition, even the interference view of forgetting is not incompatible with ACT-R's premise that forgetting is instrumental in the organism meeting the informational demands posed by the environment. For instance, it is reasonable to assume that the rate at which new associations are strengthened influences the relative accessibility of older and newer memories. Specifically, to the extent that associations between cues and new memories are strengthened, the bonds between these cues and older memories will be weakened. Such a contingency leaves open the possibility that, over ontogenetic or phylogenetic times, the rate at which new associations are strengthened is set so as to tune the accessibility of older and newer memories to the informational demands posed by the environment.

[4]For those who doubt the strictly serial nature of this search, Anderson and Lebiere (1998) have implemented a parallel connectionist model that yields the same result.

[5] Alternatively, since τ and d trade off, we could have kept d constant and varied τ, but here we do not undertake an extensive analysis of the effect of this parameter on performance. In the general discussion, however, we do consider a model in which τ is set so low that all records can be retrieved.