Fachbereich Erziehungswissenschaft und Psychologie
der Freien Universität Berlin

The Mnemonic Decision Maker:
How Search in Memory Shapes Decision Making

Dissertation
zur Erlangung des akademischen Grades
Doktor der Philosophie
(Dr. phil.)

vorgelegt von
Dipl.-Psych.
Wolfgang Gaissmaier
Berlin, 18.6.2007

Erstgutachter: Prof. Dr. Gerd Gigerenzer
Zweitgutachter: Prof. Dr. Arthur Jacobs

Acknowledgements

This dissertation is the result of research I have carried out at the Center for Adaptive Behavior and Cognition (ABC) of the Max Planck Institute for Human Development (MPI).

First of all, I dearly thank Lael Schooler who has been my main advisor in all those years. I want to thank him for all the patience he has had with me, for his intellectual, but also for his emotional support. Lael has been a great teacher and inspiring colleague, and I have learnt lots about science from him, ranging from "what is a fruitful question in the first place?" to "how can you convey a good story based on the trillions of data that you collected?" (At least I hope to have improved in those regards.) I am very grateful that he always had an open ear and helped me out with a good idea when I was stuck.

I thank Gerd Gigerenzer for creating the world's most inspiring research environment here at the MPI, and, particularly, for giving me the opportunity to work in it. Gerd is a great example of a liberal director who always supports his students and gives them the chance to take over responsibility and develop their own ideas. I thank him for all the trustful and stimulating cooperation we have had.

Jörg Rieskamp has shown me lots about modeling, parameter searches and maximum likelihood estimation, and I really want to thank him for that. Without him, I would not be able to realize all those things and other ideas in Matlab, to which Jörg has introduced me and which is a useful tool for almost everything.

I am very glad that Arndt Bröder, a former critic of ABC who now has at least partially converted, has joined this group for several months. During that time, our cooperation on the work for Chapter 2 has started, and I hope that this was only the beginning of a series of fruitful collaborations. I thank Arndt for the inspiring discussions, sometimes accompanied by a beer or two, and I also thank him for introducing me to the Scotch Malt Whisky Society.

Without Gregor Caregnato, many of my experimental dreams would not have come true. He has taken care of recruiting participants and of taking them through the experiments alive, and I really thank him for that. Also, I need to thank Christian Elsner without whom I probably would have thrown my computer out of the window at some point.

I also want to thank all the other people who have commented on earlier drafts of the chapters (or on papers based on them): Richard Anderson, Tom Beckers, Rainer Bösel, Juliet Conlin (thanks for your outstanding last minute help!), Michael Doherty, John Hutchinson, Tim Johnson, Konstantinos Katsikopoulos, Julian Marewski, Ben Newell, Robert Nosofsky, Henrik Olsson, Thorsten Pachur and Magnus Persson.

Furthermore, my thanks go to Yakoov Kareev for providing his data, Dan Bothell, Niels Taatgen and John Anderson for helping with modeling ACT-R, and Michael Kane, Richard Heitz and George Wolford for providing task materials.

Naturally, all my other colleagues have also provided very inspiring feedback on my work. They have also largely contributed to making the time I have had at ABC a great time, and we also had (and will have, I hope) a great time on conferences together. I am glad that not all the time we spent together was all too scientific!

On a more personal note, I want to dearly thank my family, my friends and Johanna, my love. How could I have survived this exciting, yet sometimes exhausting time without you?

Wolfgang Gaissmaier

Table of Contents

Introduction and Overview

About 50 years ago, Miller (1956) concluded that people can consider about seven items or categories simultaneously (plus or minus two). The limitations of human cognition concern psychologists until today. In the view of Kahneman, Slovic, and Tversky (1982), "cognitive psychology is concerned with internal processes, mental limitations, and the way in which the processes are shaped by the limitations" (p. xii). Congruently, Cowan (2001) believes that one of the "central contributions of cognitive psychology has been to explore limitations in the human capacity to store and process information" (p.87).

The goal of my dissertation will be to explore how cognitive limitations, such as a limited memory capacity, impact on human decision making in decision situations in which people need to rely on information they retrieve from memory (i.e., inferences from memory). More specifically, I am interested in how cognitive limitations shape the search for information in memory, and how people adapt their strategies accordingly. From the perspective of ecological rationality (e.g., Gigerenzer, Todd, & the ABC research group, 1999; Gigerenzer, 2004), which guided my dissertation, successful decision strategies are anchored both in the environment and in the human mind. People can exploit the core capacities of the human mind such as recall (and potentially also its limitations), which makes decision strategies simple. And they can exploit the structure of the environment, which makes decision strategies smart. Adaptive decision making thus means that people adapt their strategies both to the structure of the environment and to the limitations of the cognitive system.

Before I describe the decision situations I will investigate, ranging from very simple decision paradigms (such as a repeated binary choice probability learning paradigm) to multi-attribute decision paradigms, I briefly want to review the more general debate on how cognitive limitations affect human behavior. Basically, the two poles of the debate are, on the one hand, a pessimistic appraisal of human cognition, regarding cognitive limitations as severe liability. On the other hand, there is the view that cognitive limitations serve important functions and thus can also be beneficial. Interestingly, these views are closely tied to different views on rationality in general, which I will also illustrate.

Cognitive Limitations – Curse or Bliss or Both?

Very often, the premise of limited cognitive capacities is directly linked to its supposed negative consequences, such as reasoning errors or poor cognitive performance (e.g., Johnson-Laird, 1983). In this view, cognitive limitations force people to abandon what would be optimal decision making. Instead, people need to rely on shortcuts, on *heuristics*, which according to a pessimistic appraisal of human cognition make people vulnerable to systematic and predictable reasoning errors (e.g., Tversky & Kahneman, 1974).

This premise stems from a view of rationality as defined by the laws of logic and probability theory (see Gigerenzer & Todd, 1999). Hammond (1996) called these criteria *coherence* criteria. In this view, rational judgment is defined by unbounded rationality, which presupposes that people have unlimited time, infinite knowledge and endless reasoning powers. The proponents of this view continue to dream Leibniz's (1677/1951) dream of a universal calculus that can replace all human reasoning. According to this view, the human being should ideally be omniscient: the more information, the more processing capacity, the better. If one follows this view, it seems natural that any limitation of the cognitive system needs to pose a problem to the organism.

Another view on the limitations of the cognitive capacities of the human mind is that those limitations – such as forgetting – may serve important functions. The most important function of memory, for example, is not simply to store all information we encounter. Its most important function is to provide us with important information in specific situations. Our memory system is organized in a way which facilitates the retrieval of information which is recent or frequent (J. R. Anderson & Schooler, 1991). This is also the information we are most likely to need. Many computer programs work in a similar way by providing instant access to the files we have most recently used, which are often the files we currently need. Similarly, if you want to remember where you have parked your car, it is quite useful to have forgotten all parking lots except of the last one. There is growing evidence also from other domains (such as language acquisition) that cognitive limits can be beneficial (for an overview, see Hertwig & Todd, 2003).

From this perspective, the impact of cognitive limitations on human decision making looks quite different. Schooler and Hertwig (2005) believe that human decision strategies (such as heuristics) "have arisen over phylogenetic or ontogenetic time to exploit the existing forgetting dynamics of memory" (p. 626). Congruently, Goldstein and

Gigerenzer (2002) "consider heuristics to be adaptive strategies that evolved in tandem with fundamental psychological mechanisms" (p. 75).

In this view, the mind is an adaptation to the environment. Researchers following this perspective (e.g., Gigerenzer et al., 1999), have abandoned the beautiful dream of a universal calculus (although some of them admit that they could be easily convinced if someone were to finally show them the calculus). Instead of searching for the universal tool that can solve all problems, they believe that humans possess a repertoire of cognitive strategies (such as heuristics) that can solve specific problems. Gigerenzer et al. called this collection of cognitive strategies the *adaptive toolbox.*

The success of these cognitive strategies is anchored both in the structure of the environment and in the core capacities of the human mind (Gigerenzer, 2004). A cognitive strategy can be simple if it exploits the evolved capacities of the human mind that through evolution or learning are highly automatized, requiring little or no effort. And, in this view, the rationality of cognitive strategies is not logical, but ecological, since a strategy is not good or bad, rational or irrational per se, but only relative to an environment. Exploiting the structure of the environment makes a cognitive strategy smart.

Therefore, the right question to ask is not *whether* heuristics are successful. The right question to ask is *when* heuristics are successful. Or, to put it more generally, the right question to ask is under what circumstances a cognitive strategy (such as a heuristic) will be successful and where it will fail. Being successful here means that a cognitive strategy is successful with regard to some outside criterion such as accuracy of prediction. Hammond (1996) called assessing how good a judgment is in comparison to some outside criterion to apply a *correspondence* criterion. In the view of Herbert Simon (1990), to evaluate a cognitive strategy, one needs to consider the match between this strategy and the environment in which it operates in: "Human rational behavior is shaped by a scissors whose two blades are the structure of task environments and the computational capabilities of the actor" (p. 7). To understand human behavior, it does not suffice to look at one blade alone, such as one would not understand how scissors cut when only looking at one blade. To understand human behavior, one needs to study how the two blades fit.

A good example for the interplay between the structure of the environment and the core capacities of the human mind is the *recognition heuristic* (Goldstein & Gigerenzer, 2002). In short, the recognition heuristic uses the information whether an object is recognized or not to make inferences about some criterion value of this object. The

recognition heuristic is simple because it can rely on the human core capacity of recognition memory, which was shown to be astonishingly accurate (e.g., Shepard, 1967; Standing, 1973). Note that this does not mean that the process of recognition is simple per se; it is only simple given human recognition memory and might be tremendously hard to be implemented in a robot. The recognition heuristic will be successful in environments in which the probability of recognizing objects is correlated with the criterion to be inferred. This is, for example, the case in many geographical domains such as city or mountain size (Goldstein & Gigerenzer, 2002) and in many competitive domains such as predicting the success of tennis players (Serwe & Frings, 2006) or of political parties (Marewski, Gaissmaier, Dieckmann, Schooler, & Gigerenzer, 2005). One reason why objects with larger criterion values are more often recognized is that they are more often mentioned in the media.

The example of the recognition heuristic can illustrate that certain premises that are often made by proponents of "classical" rationality do not always hold, such as the premise that more information is always better. From the perspective of ecological rationality, more information is not better per se, but only given a certain structure of the environment; sometimes, more information may even be detrimental.

To be successful the recognition heuristic requires that an organism is partially ignorant. If an organism knows too little and recognizes none of the objects, recognition is not informative because it does not discriminate between the objects. The same happens, however, if the organism knows too much and recognizes all of the objects. Then, recognition also does not discriminate and is thereby a useless piece of information. Therefore, it can happen that people who recognize fewer objects (and thus have less knowledge) than others can make more successful predictions because they can use the recognition information more often – a less is more effect (for the exact circumstances when this can happen, see Goldstein & Gigerenzer, 2002).

The recognition heuristic may also serve as a further illustration that forgetting can be useful. Schooler and Hertwig (2005) showed that some forgetting could fuel the success of the recognition heuristic because it helps remaining the important level of partial ignorance. Without forgetting, the organism would, over time, recognize all of the objects. Then, recognition is not a useful piece of information anymore because it does not discriminate between objects. If, on the other hand, there was too much forgetting, the organism would not be able to recognize any of the objects, leaving it again puzzled. The

key to success lies in recognizing some, but not all of the objects, and forgetting helps to keep it that way.

To answer the question posed in the title of this subsection – whether cognitive limitations are curse or bliss or both – I would argue that cognitive limitations are neither curse nor bliss, they can be both. I strongly believe that cognitive limitations are functional with regard to certain tasks in certain situations. This, however, may come with a price in other situations. Thus, the answer is an answer that every journalist interviewing a scientist hates: It depends on the situation. In my dissertation, I investigated how limitations of the memory system affect decision making in different decision situations, in particular with regard to how information is searched in memory. In the next section, I will give a brief overview of those different decision situations, and thereby of the different chapters.

Outline of the Dissertation

The question that ties all the chapters together is the question how people search for information in memory when they make decisions. This information search in memory is shaped by the limitations of the cognitive system, which will constrain people, for example, in which order they search for information or how much information they are able to search for. The idea here is that these constrains may be functional because they facilitate specific ways of searching for information, which are adaptive in certain situations.

An important distinction in all of the chapters is the distinction between *understanding* and *evaluating* cognitive processes. I believe that somewhat artificial experiments which do not represent the outside world are often needed to understand and describe cognitive processes. In some sense, this is similar to studying how the visual system works with stimuli that are deliberately designed to trick the system, resulting in visual illusions. They help us to understand how the visual system works. To evaluate processes, visual or cognitive ones, however, it is important to consider how such a process would fare in the outside world. This distinction draws on concepts such as ecological rationality (see above) and on Brunswik's idea of representative design (e.g., Brunswik, 1956), in which he stresses that a generalization of results requires that variables are representative of a carefully defined set of conditions. These important issues will reoccur at several points throughout the dissertation. In this regard, please also note

that the chapters were written to be understandable on their own. Therefore, there may be some redundancies especially with regard to these general issues.

The work in Chapter 1 was inspired by the stunning result that a lower short-term memory capacity benefits performance on a correlation detection task (Kareev, Lieberman, & Lev, 1997). In the task people successively encountered envelopes with two different colors and each time had to decide which out of two objects they think they will contain. Kareev et al. assumed that low spans perceived the correlations as more extreme because they relied on smaller samples, and small samples were shown to overestimate the correlation coefficient. However, this *small sample hypothesis* has been criticized recently, because small samples also bear a higher risk of false alarms (Juslin & Olsson, 2005; R. B. Anderson, Doherty, Berg, & Friedrich, 2005).

Therefore, I put forward an alternative explanation that draws on the idea that low spans are more strongly constrained in their memory search and will adapt their strategies accordingly. Instead of assuming that low spans compute the correlation coefficient based on smaller samples, however, I assume that they explore the space of hypotheses how to improve performance on the task to a lesser degree. A typical hypothesis people have in those tasks is that there is a pattern in the sequence of events. But since there is no pattern in the sequence, such an exploration is counterproductive. Therefore, people who explore less, fuelled for example by capacity limitations, could be more successful.

However, less exploration could come with the price that it makes it harder to adapt to an environment which is changing. This idea will be modeled precisely in the cognitive architecture ACT-R (e.g., J. R. Anderson & Lebiere, 1998) and investigated experimentally. The results will be discussed in terms of ecological rationality and representative design. Explorative behavior cannot be finally evaluated by considering its success in those experiments. Instead, it needs to be considered in which environmental structures outside of the laboratory it will be successful.

In Chapter 2, I will study more complex situations, in which people face several attributes (cues) which they could use when making decisions. This chapter was inspired by the work of Bröder and Schiffer (2003b, 2006). In contrast to the usual screen-based research paradigm, Bröder and Schiffer implemented the idea of memory search in cue based decisions by introducing a cue-learning paradigm in which participants acquired knowledge about cues describing objects. These cues then had to be retrieved from memory when making decisions between objects in a decision phase. In sum, the results

by Bröder and Schiffer show that the need to retrieve cue information from memory induced the use of simple decision strategies, especially when cues were represented verbally and when working memory load was high. These results link to the idea explored in Chapter 1 that the search for information in memory is constrained by cognitive limitations and that people adapt their strategies to demands imposed on their memory system.

The decision strategies Bröder and Schiffer (2003b, 2006) investigated rest on the strong assumption that people should process information retrieved from memory in a sequential manner. A methodological challenge in studying memory-based decisions is, however, that the search process is not directly observable, so that Bröder and Schiffer determined which strategy someone apparently used solely on the outcomes of the decisions. And such an outcome-based strategy classification remains ambiguous: A simultaneous global matching process with cue weights chosen appropriately could mimic the decision outcomes predicted by sequential search strategies, albeit assuming cognitive processes that are very different. Therefore, I have reanalyzed response times from Bröder and Schiffer's published experiments and of one new experiment, which could serve as convergent evidence for the assumption of sequential search.

The main goal of Chapter 2 was to understand the process underlying memory-based multi-attribute decisions. It was assumed that people often rely on simple cue-based strategies that process information sequentially in memory-based decisions. Chapter 2 remained silent on the evaluation of those strategies. This is where Chapter 2 is complemented by Chapter 3. In Chapter 3, I want to explore the possibility that people could exploit their memory system in a way that facilitates sequential search strategies and at the same time is successful in a real world environment. The success of sequential search strategies depends crucially on the way in which cues are ordered. Juslin and Persson (2002) argued that some of the heuristics introduced by Gigerenzer et al. (1999) and studied in Chapter 2 presuppose quite some knowledge about how to order cues and thus are not so simple after all. Therefore, I investigated a very simple way to order cues which exploits central features of the memory system. In particular, I addressed the question whether the speed of retrieving information (i.e., the *fluency* of information) could be used successfully to order cues.

The speed of retrieving information is mostly a function of frequency and recency of encountering this information (J. R. Anderson & Schooler, 1991), and also the context

of the information is important (Schooler & Anderson, 1997). That is, people are most likely to come up with cues quickly they have encountered frequently and recently. There could thus be some peculiarities of the environment which betting on the speed of retrieval might exploit. For example, positive cue values (i.e., cue values indicating the presence of a cue) should be more fluently available for large objects, since people will encounter information about those large objects more frequently in the environment. For negative cue values, this relation might be the exact opposite. That is, the fluency of negative cue values could be negatively related to city size (i.e., retrieved more quickly for smaller cities), which again is informative. A further feature of the memory system a retrieval based strategy could exploit is that correct cue knowledge should come to mind more quickly than incorrect cue knowledge, which is a common finding in the memory literature (e.g., Ratcliff & Smith, 2004).

To investigate whether those exploitable features do occur, an experiment has been conducted to assess the fluencies of stimuli that people have encountered in the real world outside of the laboratory – German cities described on several attributes such as whether a city has an airport, a soccer team, etc. These fluencies were then used to actually simulate the performance of strategies using the fluency of cues to order information. The success of those strategies was compared to other decision models such as heuristics described by Gigerenzer et al. (1999) and multiple regression.

1 Chapter 1

Simple Predictions Fueled by Capacity Limitations: When are They Successful?

In this chapter, I discuss the hypothesis forwarded by Kareev and colleagues that a limited short-term memory capacity fosters the detection of correlations (Kareev, 1995a, 1995b, 2000, 2004; Kareev, Lieberman, & Lev, 1997). They made the counterintuitive prediction that limited capacities are beneficial in correlation detection because they force people to rely on small samples. This prediction was derived from the statistical fact that correlations tend to be overestimated in small samples, which was initially supported by behavioral data. However, their theoretical account has been challenged recently because small samples also yield a higher risk of false alarms (R. B. Anderson, Doherty, Berg, & Friedrich, 2005; Juslin & Olsson, 2005). Furthermore, I will review empirical evidence that is in conflict with Kareev's theoretical account. These challenges call for the exploration of alternative explanations for the findings that Kareev and colleagues interpreted as supporting their theory. My alternative explanation, drawn from the probability learning literature, was tested against Kareev's hypothesis. Before reporting these results, I describe the domain of correlation detection and explain Kareev's arguments and their challenges in more detail.

1.1 Limited Capacities and Correlation Detection: The Small Sample Hypothesis

Correlation detection (or, more generally, contingency assessment) is considered to be an important component of adaptive behavior, and has been studied in a variety of domains and with a variety of tasks (for reviews, see Alloy & Tabachnik, 1984; De Houwer & Beckers, 2002). Most studies of contingency assessment are concerned with contingencies between binary variables. They can be described by a two-by-two contingency table (see Figure 1.1) that shows the frequencies (or probabilities) of the presence or absence of one variable (outcome, e.g., a disease), given the presence or absence of another variable (input, e.g., a symptom).

	Outcome1	Outcome2	
Input1	a	b	a + b
Input2	c	d	c + d
	a + c	b + d	a + b + c + d
			= *N*

Figure 1.1. Prototypical contingency table.

The phi coefficient[1], a common measure to compute contingencies between binary variables, is defined as

$$\Phi = (ad - bc) / \sqrt{(a+b)(c+d)(a+c)(b+d)} \,. \qquad (1)$$

Kareev (1995b) argued that people rely on samples from the environment to assess correlations between, for example, two dimensions of a set of objects. The size of these samples is supposed to be bounded by short-term memory capacity. In a theoretical analysis, Kareev concluded that the use of small sample sizes facilitates the early detection of correlations by amplifying them. Specifically, both the median and the mode of the sampling distribution exceed the population correlation, and the smaller the sample, the more so. Building on the assumption that people's perception of correlation is the result of calculating the correlation on the basis of a sample, Kareev assumed that consideration of a small sample is more likely to result in a more extreme perception of correlation. Since the samples people consider are smaller for people with a lower short-term memory capacity (low spans) than for those with a higher short-term memory capacity (high spans), the argument goes, low spans should be more likely to perceive the correlation as more extreme, and thereby detect it earlier.

Kareev and his colleagues provided experimental support for this theoretical argument since low spans indeed performed better on a correlation detection task (Kareev et al., 1997). The task consisted of predicting, trial-by-trial, which of two possible symbols (X or O) an envelope (which could be either red or green) contained. The number of Xs and Os within the envelopes was varied to yield correlations ranging from $\Phi = -.60$ to $\Phi =$

.60. A correlation here means that, for example, there are more Xs in red envelopes and more Os in green envelopes. Detecting this correlation helps people to increase their predictive performance. I will refer to this task as the envelope task. Kareev et al. concluded that people with a lower short-term memory capacity, and hence a smaller sample size to consider, "perceived the correlation as more extreme and were more accurate in their predictions" (p. 278). I will call this Kareev's *small sample hypothesis* of correlation detection in the remainder of this chapter.

The phenomenon of a low capacity advantage in correlation detection is particularly surprising, considering that short-term memory capacity has generally been found to be positively correlated with a variety of cognitive abilities, for example, executive functioning (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001) or performance on the scholastic aptitude test (SAT, Engle, Tuholsky, Laughlin, & Conway, 1999). The correlation between the related construct of working memory capacity and reasoning ability is even more pronounced (Kyllonen & Christal, 1990). Moreover, the theoretical explanation of this low capacity advantage, the small sample hypothesis, has been criticized on theoretical grounds, and there is also conflicting empirical evidence, both of which will be reviewed in the following.

1.1.1 Theoretical limitations of the small sample hypothesis

Juslin and Olsson (2005) criticized Kareev (2000) for only taking into account the hit rate when discussing the small sample advantage, that is, detecting a sample correlation (Φ) given that there is a population correlation (ρ), p(Φ|ρ). In contrast, Juslin and Olsson stressed the importance of the posterior probability of a hit, p(ρ|Φ), which takes into account false alarms (i.e., believing that there is a positive correlation when it is in fact zero or negative). Juslin and Olsson's analyses of considering how likely it is that one correctly infers that there is a population correlation ρ based on a sample correlation Φ lead to the conclusion that the alleged benefits of small samples do not occur.

---

[1] If correlations are symmetrical (i.e., $a - b = d - c$) and marginal distributions are equal (i.e., $a + b = c + d$), the phi coefficient leads to the same nominal value as ΔP, defined as $\Delta P = a/(a + b) - c/(c + d)$.

R. B. Anderson et al. (2005), using a signal detection approach, specified some conditions under which a small sample advantage could hold, even if one takes false alarms into account. Their simulations demonstrated that a small sample advantage can exist if one makes the additional assumption that people only decide that a correlation is present in the population when the correlation they observe in the sample exceeds a decision threshold. Otherwise, the observed correlation is ignored. If the decision threshold is above or equal to the correlation in the population, a small sample advantage exists. For more liberal correlation thresholds (i.e., between zero and the population parameter), however, there is a large sample advantage.

In response to these criticisms, Kareev refined the small sample hypothesis, arguing that a small sample advantage is only possible for large correlations (Kareev, 2005, see also Kareev, 2000). However, this restriction makes it problematic to explain, with the small sample hypothesis, the low capacity advantage observed in Kareev et al. (1997, Experiment 1) because a low capacity advantage was also observed for small correlations. Moreover, additional empirical evidence conflicting with the small sample hypothesis also exists, which will be reviewed in the following.

1.1.2 Conflicting empirical evidence

Kareev et al. (1997) assumed that people who consider smaller samples are likely to perceive correlations as more extreme than they actually are in the population. From this assumption, it follows that people should also estimate correlations as being higher when they base their estimate on a small, compared to a large, sample. However, in experiments in which participants repeatedly explicitly estimate correlations, participants do not estimate higher correlations based on smaller samples, but rather the tendency is that those estimates increase with increasing sample size (e.g., Clément, Mercier, & Pasto, 2002; Shanks, 1985, 1987). Moreover, studies with measures related to short-term memory capacity suggest that people with lower capacities are less accurate in correlation assessment. For instance, such people include those with lower general cognitive ability, measured by SAT scores (Stanovich & West, 1998), who are elderly (e.g., Mutter & Williams, 2004; Parr & Mercier, 1998), and who are performing under increased memory demands (Shaklee & Mims, 1982). That is, in correlation assessment, neither a small sample nor a low capacity advantage has been reported, but rather the opposite. How then can the empirical finding of the low capacity advantage on correlation detection reported by Kareev et al. (1997) be reconciled with these other results?

1.2 An Alternative Explanation: Differences in Predictive Behavior

Juslin and Olsson's (2005) arguments imply that Kareev et al.'s (1997) task is not really about the detection of correlation. Participants did not have to detect a correlation among trials with a correlation present (signal trials) and trials without a correlation (noise trials), but they were separately tested on either signal or noise trials, thereby not encountering the risk of false alarms. Since the task therefore does not really pose a detection problem, one cannot conclusively argue for a low capacity advantage in the detection of correlation. It has only been shown that low spans are more successful given that there is a correlation. The theoretical limitations of the small sample hypothesis suggest that a different cognitive mechanism could underlie this low capacity advantage. In the following, I will develop an alternative explanation which depends on reinterpreting the task as simple probability learning.

Kareev et al. (1997) assume that a low capacity advantage in correlation detection stems from a more exaggerated perception of correlation. However, the envelope task (Kareev et al., Experiment 1) did not assess differences in the perception of correlation. Kareev et al. refrained from asking their participants about their perception, but rather inferred their perception from their predictive behavior (a term used by Estes, 1976, for example). That is, they counted how often a participant predicted an event, given the color of the envelope, for example, how often he or she predicted X, given a red envelope. These frequencies were used to compute what Kareev et al. called the perceived correlation by entering them into a contingency table, such as Figure 1, and determining the phi correlation from this table. Inferring perception from behavior requires the strong assumption that people predict events exactly with the relative frequency with which they perceive them.

In my view, it is necessary to disentangle perception and predictive behavior because predictive behavior can differ between people who perceive the same correlation. Thus, it is possible that differences in predictive behavior alone could be sufficient to explain the low capacity advantage. To understand the difference between perception and predictive behavior, and to understand how differences in predictive behavior could be related to capacity limitations, I next draw a connection to the probability learning literature that reaches back to Brunswik (1939) and Humphreys (1939), and which has been extensively studied since the 1950s (e.g., Estes & Straughan, 1954; for reviews, see Myers, 1976; Vulkan, 2000).

1.2.1 Correlation detection as probability learning

The typical probability learning task consists of repeatedly predicting which of two events will occur next, with one event usually having a higher probability of occurrence. The correlation detection task used by Kareev et al. (1997) is similar because it also requires predicting one out of two events (the symbols X and O, given the color of the envelope). Bauer (1972), for example, used a task that is almost identical to the one used by Kareev et al. However, she did not cast it as a correlation detection task, but rather as a probability learning task with two cues (the colors) and criterion events (the symbols).

A very simple predictive behavior that performs well is to always predict the event that, so far, has been observed most frequently. For example, if one event occurs with a probability of 70%, always predicting this event will result in an accuracy of 70%, on average. This behavior is called maximizing. Most often, it has been found, however, that the majority of people do not maximize. Instead, what is often found is probability matching (Vulkan, 2000), which consists of predicting an event in proportion to its probability of occurrence (i.e., an event that occurs with a probability of 70% is predicted to occur in 70% of the trials). Probability matching, on average, leads to lower accuracy (i.e., expected accuracy of .7*.7+.3*.3 = .58).

The distinction between maximizing and probability matching is relevant for the correlation detection task used by Kareev et al. (1997). Consider the two types of envelopes and the conditional probabilities of the events, given the color of the envelope. Maximizing implies always predicting X when, for example, a red envelope is shown, if X has been observed more frequently in the past when opening red envelopes. Given a correlation between the envelopes' color and the symbols, this would then imply always predicting O when a green envelope is encountered[2].

Two people may share a perfect perception of the probabilities of the events (or of the correlation), but behave differently, for instance, by probability matching or maximizing. In contrast, Kareev et al.'s (1997) assumption that it is possible to deduce perception from behavior presupposes that everyone's behavior matches their perception

[2] However, it is interesting to note that in the case of asymmetric marginal distributions, it can occur that even when observing a positive correlation between two variables, maximizing implies always predicting the same event. For example, consider a sample of 20 red and 20 green envelopes. Imagine that 19 Xs had been observed in red envelopes and 11 Xs had been observed in green envelopes, which leads to a substantial phi coefficient of .46. Nevertheless, since X is the most frequent event for both kinds of envelopes, maximizing implies predicting X every time.

of the conditional probabilities, but that low spans have a distorted perception of these probabilities.

Moreover, Kareev et al.'s (1997) explanation requires that people actually think about the task in terms of the correlation between the color of the envelopes and the frequencies of the different symbols within them. But just because this task can be described as a correlation detection task does not mean that the participants view it this way. From the probability learning perspective (e.g., Bauer, 1972) one could assume that participants learn the conditional probabilities of a symbol given a color independently for both colors. Thus, the perception of correlation argument would not be applicable. But then, I need to explain how short-term memory limitations could be beneficial from the probability learning perspective, which I will do next.

1.2.2 Maximizing is fostered by limited memory capacities

The probability learning literature has struggled with the phenomenon of probability matching because it seems inconsistent with a person's goal to maximize his or her payoff. West and Stanovich (2003) argued that this inconsistency results from insufficient cognitive capabilities, and it has been shown that this inconsistency can be reduced with extensive training and high monetary payoffs (e.g., Shanks, Tunney, & McCarthy, 2002). At odds with this perspective that people are not smart enough to maximize, is evidence that reduced or limited memory capacities are associated with a higher prevalence of maximizing.

On the one hand, there are studies demonstrating that people with lower memory capacities maximize more frequently. Maximizing was shown to be more prevalent for people with lower intellectual abilities (Singer, 1967), for children (Derks & Paclisanu, 1967; Weir, 1964), and for different kinds of animals, such as pigeons (Herrnstein & Loveland, 1975, Hinson & Staddon, 1983), rats (Bitterman, Wodinsky, & Candland, 1958), and monkeys (Wilson & Rollin, 1959). On the other hand, the likelihood of maximizing is higher for people under the cognitive load of a secondary task, which was shown with a concurrent estimation task (Bauer, 1972; Neimark & Shuford, 1959) and with a verbal working memory task (Wolford, Newman, Miller, & Wig, 2004).

An explanation for this could be that maximizing is very simple – a feature that is often overlooked (Bauer, 1972). In contrast, probability matching could be the remnant of more involved cognitive processes, such as searching for patterns in the sequence of events, which has been nicely demonstrated in a probability learning study by Yellott

(1969). In the last block of his experiment, participants always received feedback indicating that their predictions were correct, irrespective of what they predicted. They continued to match probabilities as they did previously, and when they were asked for their impressions afterwards, most responded that they finally found the pattern in the sequence. Wolford, Miller and Gazzaniga (2000) hypothesized that the search for such a pattern necessarily results in behavior that appears to be probability matching because every reasonable pattern will tend to match the probabilities.

Preventing complex hypothesis testing, such as searching for patterns by means of instruction, for example, by telling people that the best they could do is reaching an accuracy of 75% (Fantino & Esfandiari, 2002), or by making the task look like a gambling task and not a problem solving task (Goodnow, 1955), increased the prevalence of maximizing. Since working memory capacity is related to hypothesis generation (Dougherty & Hunter, 2003), lower memory capacities could foster maximizing by making complex hypothesis testing, and thereby complex predictive behavior, less likely because it is more memory demanding.

1.3 Summary: Differences in Perception Versus Differences in Predictive Behavior

The findings that people with lower or reduced memory capacities show a higher prevalence of maximizing could present a plausible alternative explanation for the low capacity advantage found by Kareev et al. (1997). This implies that low spans are more likely to maximize because they are less likely to test complex hypotheses, and are thereby more likely to settle on simple maximizing. The reasoning behind this explanation and the explanation given by Kareev et al. is strikingly different. Kareev et al. stressed the influence of short-term memory capacity on the perception of correlation, which implies that the behavioral response to the perception is always identical, while my alternative explanation builds on the idea that people could very well share the same accurate perception, but still differ in how they respond to their perception. That is, I have here two competing hypotheses, Kareev et al.'s *small sample hypothesis* and my alternative explanation, the *predictive behavior hypothesis*.

1.4 Modeling the Competing Hypotheses in ACT-R

The central goal of this chapter consists of pitting these hypotheses for a low capacity advantage on the correlation detection task, used by Kareev et al. (1997), against each other. An important step in doing this, which Kareev has not yet carried out, is to specify a precise computational models of the cognitive process underlying correlation detection. I think it is important that the model specifies the learning process, resulting in a certain perception of correlation and the behavioral response, so that both processes can be disentangled. To model the processes, I use ACT-R which has been developed by Anderson and his colleagues (e.g., J. R. Anderson & Lebiere, 1998; J. R. Anderson et al., 2004). ACT-R models are able to account for a wide variety of phenomena including, for example, practice and retention (J. R. Anderson, Fincham, & Douglass, 1999), decision making (Gonzalez, Lerch, & Lebiere, 2003), language learning (Taatgen & Anderson, 2002), and, important for me, probability learning (Lovett, 1998). Implementing the correlation detection task in ACT-R allows me to model the explanation for a low capacity advantage on the basis of differences in perception, as provided by Kareev et al. (1997), versus the explanation based on differences in predictive behavior. Thereby, these models allow me to make divergent predictions for people who differ in their short-term memory capacity.

1.4.1 Implementing the correlation detection task in ACT-R

The core of ACT-R is constituted by the declarative memory system for facts and the procedural system for rules. Here, I focus on the declarative memory system to model the correlation detection task which results in an instance-based model, building on Logan's (1988) idea that previous solutions to a problem are stored in memory as examples that can be retrieved to solve future problems (for a more detailed description of instance learning in ACT-R, see Taatgen, Lebiere, & Anderson, 2006). The declarative memory system consists of chunks that represent information (e.g., about the outside world, about oneself, about possible actions, etc.). These chunks take on activations that determine their accessibility. That is, whether they can be retrieved. When applied to the correlation detection task, chunks represent instances of possible responses to the envelopes encountered in each trial. Altogether, there are four chunks to represent all possible combinations of the envelopes' two colors and the two possible events connected

with the envelopes (i.e., “red envelope: X”, “red envelope: O”, “green envelope: X”, and “green envelope: O”). As a consequence of following ACT-R’s standard rule for reinforcing chunks, the history of how often and when chunks have been used in the past determines their activation (see below). Since activation is a combination of frequency and recency, different histories can lead to the same activation at any given moment of time.

The model represents the cognitive processes of one single individual solving the envelope task. Each time an envelope is presented, the model attempts to retrieve one of the two responses associated with the envelope’s color. For example, if there is a red envelope, the model attempts to retrieve the chunks “red X” and “red O.” These two chunks enter a retrieval competition since only one of them can be retrieved at a time. The likelihood of each chunk winning this competition depends on their activations. The more frequently and the more recently a chunk has been used the higher its activation. The combination of the chunks’ activation levels determines the probability that any one chunk will be retrieved and so determines the model’s predicted response. After the response, the model receives feedback whether it was right or wrong, which leads to reinforcing the chunk representing the correct answer. Thus, the chunk that was retrieved and triggered the response, and the correct chunk, are reinforced, which thereby strengthens their activation. This also implies that a correct answer will be reinforced twice, while an incorrect answer results in reinforcing both the chosen and the correct response once.

Formal definitions. Formally, the activation of a certain chunk *i* is defined as

$$A_i = B_i + \sum_j w_j S_{ji} \qquad (2)$$

where B_i is the base-level activation of chunk i that reflects its learning history, the *Wjs* reflect the attentional weighting of the elements that are part of the current goal, and the *Sjis* are the strengths of association from the elements *j* of the current context to chunk *i*. For my purpose, only the base-level activation is relevant. The base-level activation of a chunk is defined by

$$B_i = \ln(\sum_{j=1}^{n} t_j^{-d}) \qquad (3)$$

where t_j is the time since the *j*-th practice of an item and *d* is a decay parameter for which .5 has emerged as a default value across a variety of studies (J. R. Anderson et al., 2004). A chunk can only be retrieved if its activation Ai is above a retrieval threshold τ, accordingly the probability that a chunk is retrieved is

$$P_i = \frac{1}{1 + e^{-(A_i - \tau)/\sqrt{2}s}} \quad (4)$$

where s controls the noise of the retrieval process. If there is more than one chunk that matches a retrieval request, as there is here, the probability that a particular chunk is retrieved is

$$P_i = \frac{e^{A_i/\sqrt{2}s}}{\sum_k e^{Ak/\sqrt{2}s}} \quad (5)$$

If a chunk has been retrieved, the retrieval time is defined as

$$T_i = Fe^{-A_i} \quad (6)$$

where F is a latency factor.

Parameters that relate to the competing hypotheses. There are two parameters that are of interest to me since they can be related to the two hypotheses (small sample hypothesis vs. predictive behavior hypothesis), the decay parameter d in the base-level learning equation and the noise parameter s in the equation specifying the probability of winning the retrieval competition. The decay parameter d affects the impact of recency on the activation of chunks. Note that there is no differentiation between short- and long-term memory in ACT-R. The base-level learning equation which produces rapid initial decay and slower later decay is key to accounts of both short-term memory tasks, such as memory span, and long-term memory tasks, such as free recall (J. R. Anderson, Bothell, Lebiere, & Matessa, 1998). Without decay, each outcome would be weighed equally, irrespective of how long ago it has been observed. A model with high decay puts more weight on recent information and tends to disregard old information. I believe that this parameter offers a precise way to relate the small sample hypothesis proposed by Kareev (1995b; Kareev et al., 1997) to processes in ACT-R. The higher the impact of recency, the fewer items are important for a decision, which leads to paying attention to a small sample.

The noise parameter s affects how likely it is that the more activated chunk will win the competition. Without noise (i.e., $s = 0$), the most activated chunk will always be retrieved (given that it is above the retrieval threshold τ), resulting in perfect maximizing in the limit. Higher noise allows less activated chunks to be retrieved from time to time. While such noise results in suboptimal behavior under some conditions, it is also used to model exploration (Taatgen et al., 2006). Thus, the noise parameter provides a simple way to model facets of predictive behavior, without developing a precise model of how people

go about searching for patterns. In this regard, it is important not to interpret noise solely as error. Rather, higher levels of noise capture a proliferation of hypotheses that a participant may entertain, yielding behavior that looks like the model is searching for patterns in the data. This searching results in probability matching (the precise value for *s* leading to probability matching behavior depends on the task), whereas low levels of noise result in deterministic maximizing behavior. I think that the higher complexity of this behavior makes the relation to short-term memory plausible. Therefore, I believe that variation in this parameter nicely captures the predictive behavior hypothesis.

1.4.2 Methods

I used two variants of the model to instantiate the two hypotheses for explaining the low capacity advantage. With the first decay variant of the model, I represent Kareev's small sample hypothesis, with fast decay resulting in a focus on a small sample of recent events. With the second noise variant of the model, I represent the predictive behavior hypothesis, with low noise resulting in deterministic maximizing behavior.

Kareev kindly provided me with the data from Kareev et al.'s (1997) Experiment 1, with which I constrained the models used in my simulations. I chose the 128 trials from the conditions with $\Phi = |.375|$ with symmetric distributions of Xs and Os contained in the envelopes (i.e., there were 44 Xs [68.75%] and 20 Os [31.25%] contained in envelopes of one color, while this was exactly reversed for the other color). This is the condition that I also used in my experiments (see Chapter 3). Note that the qualitative modeling results did not depend on the actual correlation, that is, modeling other conditions yielded the same qualitative results. The model was fit to the relative frequency of maximizing responses, that is, the average proportion choosing the maximizing answer on a particular trial which was further averaged within four blocks consisting of 32 trials each. This was done separately for high and low spans as defined by Kareev et al.

While it is, in principle, possible to differentiate between the two model variants quantitatively on the trials that were fitted, it is not possible to disentangle the two hypotheses qualitatively on those trials. Therefore, I considered a manipulation that distinguishes between the two model variants, and thereby the two hypotheses. A change in the correlational structure of the environment (simply referred to as shift in the following) allows for such a differentiation (see below). That is, after the initial 128 trials with a correlation of $\Phi = +.375$, I added 128 trials in which the correlation (i.e., the probability of each event given one or the other color) was exactly reversed, that is, $\Phi = –$

.375. If, for example, red was predictive for X in the 128 fitting trials, it was predictive for O in the trials after the shift. Thus, I made predictions for how high and low spans would adapt their behavior to this shift, depending on the variant of the model, and thereby the hypothesis[3]. However, note that this shift was not implemented in Kareev et al.'s (1997) experiment. Thus, I first fitted the two model variants to Kareev et al.'s data, and second, the fitted models were used to predict behavior for a hypothetical shift not conducted by Kareev et al.

To fit the models to Kareev et al.'s (1997) data, only the one parameter representing either of the hypotheses was varied in each of the model variants. That is, in the decay variant of the model, only decay *d* was varied to fit the curves of both low and high spans separately, while noise *s* was held constant. In the noise variant, only noise *s* was varied to fit the curves of both low and high spans separately, while decay *d* was held constant. All other parameters were set to identical values for both model variants. My parameter search was informal, and there is no guarantee that they produce optimal fits on Kareev et al.'s data. But I was mostly interested in the predictions made by the two model variants after the hypothetical shifts, and there the qualitative results of the model did not change within a wide range of parameter values. Each simulation was run 10,000 times to obtain reliable results.

1.4.3 Results

Given the simplicity of the task, I think it is unrealistic that people fail to retrieve an answer at any point in time. Therefore, the retrieval threshold τ was set to -10 to ensure that the model never fails to retrieve a chunk in both model variants. The latency factor *F* was set to .1. These parameter values are well within the range of parameter values commonly used (see J. R. Anderson & Lebiere, 1998). In the decay variant, I found the best fit for low spans by setting *d* either to be fast (1, representing low spans, $R^2 = .70$) or absent (0, representing high spans, $R^2 = .93$), while keeping the noise *s* constant at .5.

---

[3] For convenience, I present the theory here in its complete form, although it was formalized after running the experiments. Initially, I started out with the informal hypothesis that if a low short-term memory capacity helps people in detecting correlations, then it should also help them in detecting a change in the correlation. The predictive behavior hypothesis was developed after Experiment 1.

In the noise variant, I obtained a good fit by setting the noise *s* to either .45 for low spans (R^2 = .74) or .6 for high spans (R^2 =.95), while keeping the decay *d* constant at its default value of .5. Overall, the predictions of both model variants are quite good since both models appropriately describe the increasing frequency of maximizing. However, both models miss the drop in the relative frequency of maximizing that the low spans exhibit on the third block, which explains the lower fit for low spans (see Figure 1.2).

Before the shift, the decay variant of the model predicted a higher frequency of maximizing with a higher decay parameter value, representing faster forgetting, and thereby capturing the behavior of low spans. The noise variant of the model captures the behavior of low spans with the lower value of noise because lower noise predicts a higher frequency of maximizing, representing a more deterministic response. Therefore, both variants of the model allow for the prediction of a difference in maximizing behavior for low and high spans, although based on different mechanisms. However, the decay parameter *d* was not able to fully capture the magnitude of the gap separating the curves.

A clear difference between the predictions of the two variants of the models emerged after the shift. Faster decay also led to increased maximizing after a shift. Thus, according to the decay variant, low spans should perform better both before and after a shift. Moreover, the predicted fast decay advantage is even more pronounced after the shift than before. However, the opposite prediction was observed for the noise variant of the model. Lower noise yielded decreased and not increased maximizing after a shift. The chunks with the highest activations before the shift favor the wrong choice after the shift. Thus, it is likely that a chunk is retrieved which results in an incorrect (i.e., non-maximizing) answer after a shift, the lower the noise, the more so. Thus, according to the noise variant, high spans who did worse before a shift should outperform low spans after the shift[4]. Figure 1.2 shows the predictions of the two variants of the model.

---

[4] Note, however, that this only holds until the activation of the chunk representing the correct (i.e., maximizing) answer is strengthened enough so that it surpasses the activation of the chunk representing the wrong answer. Then, lower noise would turn out to be beneficial once more. That is, the disadvantage after a shift resulting from lower noise will only hold as long as the relative frequency of maximizing is below .5, on average. Therefore, this noise variant of the model predicts that, over time, low spans catch up with high spans, and even outperform them subsequent to many trials after the shift.

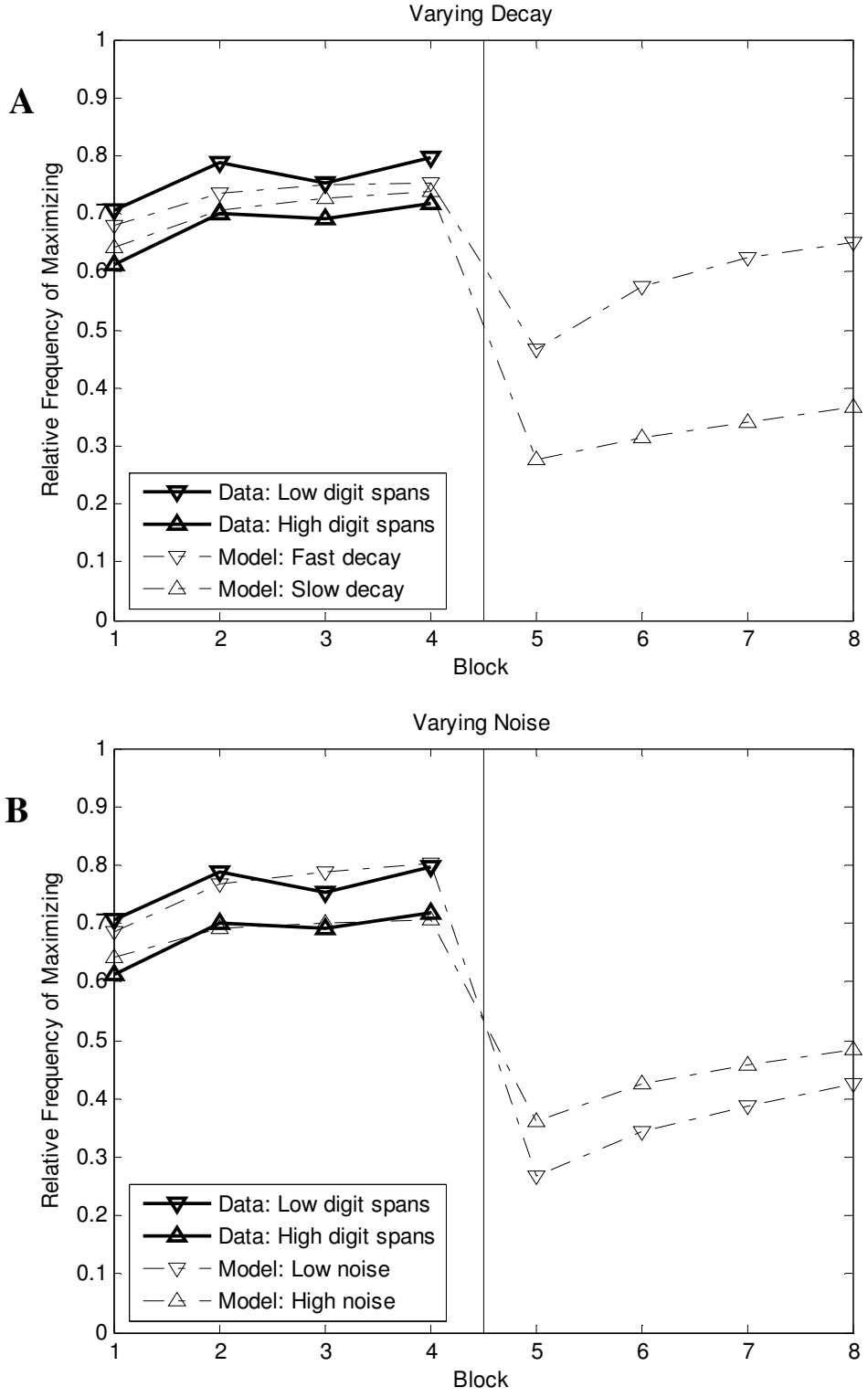

Figure 1.2. Model predictions of (A) the decay and (B) the noise variant. The models were fitted to data on 4 blocks of 32 trials each, and then predictions were made for behavior after a shift in the environment (indicated by the vertical line).

1.4.4 Is the ACT-R model equivalent to the small sample hypothesis?

One could argue that the ACT-R model accounting for the small sample hypothesis (the decay model) is not a strict translation of the small sample hypothesis, because the decay model assumes that people rely on samples biased to include more recent items, while the small sample hypothesis assumes that people rely on random samples (see, e.g., Karrev, 2004). Therefore, I tested whether strictly translating the random sample procedure of the small sample hypothesis results in qualitatively identical predictions to my decay model. That is, I wanted to find out whether small samples are, in a comparable manner, also better for detecting a shift in the environment than larger samples when using randomly selected samples.

Method. I simulated the late-shift condition of my experiment with 256 trials with a correlation of $\Phi = .375$ followed by 128 trials with a correlation of $\Phi = -.375$. Recall that there were envelopes with two different colors containing one of two different symbols. One symbol (e.g., X) had a prevalence of .6875 within one kind of envelope (e.g., red) and a prevalence of .3125 in the other (e.g., green), while this was reversed for the other symbol. After the shift, the distribution of symbols within the envelopes was exactly reversed (i.e., if X was more prevalent in red envelopes before the shift, it was more prevalent in green envelopes afterwards).

I was interested in how quickly the shift would be detected when random samples consisting of different numbers of trials were drawn from all previous trials. On each trial after the shift, I randomly sampled between 4 and 10 previous trials without replacement and computed the sample correlation to see whether it indicated a "correlation more extreme than that of the population" (Kareev et al., 1997, p. 278). That is, the sample correlation had to be more negative than $\Phi = -.375$ to count as detection, which is also consistent with R. B. Anderson et al.'s (2005) notion of a decision criterion that is necessary for a small sample advantage to exist. The simulation was run 1000 times.

Results. Smaller samples were indeed better for detecting the shift than larger samples. Figure 1.3 illustrates this result for sample sizes 4, 7, and 10.

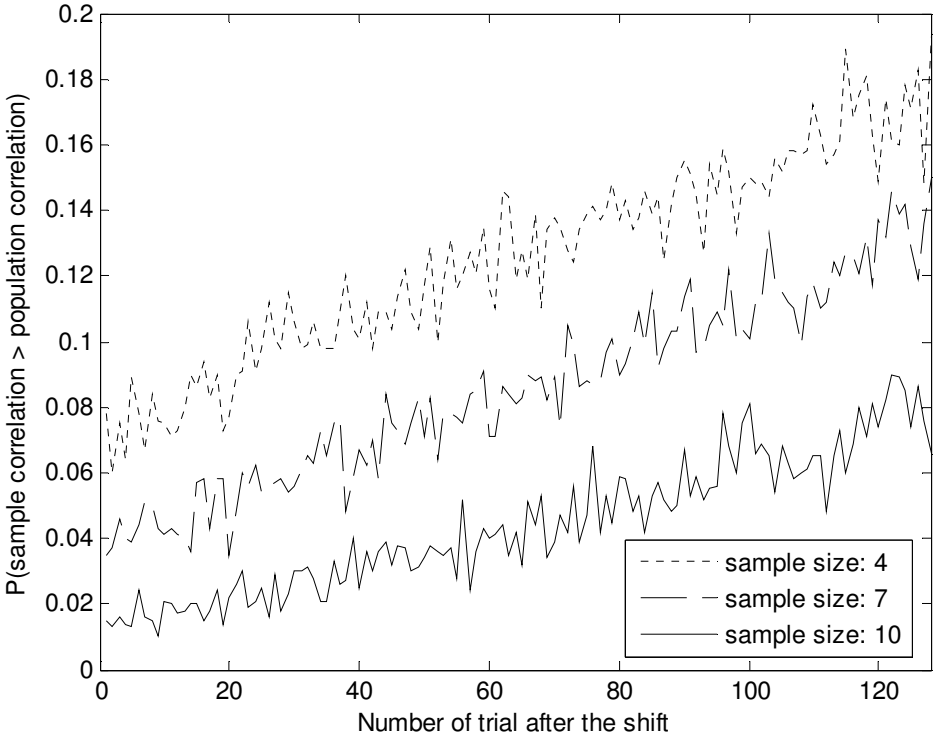

Figure 1.3. Probability of encountering a sample correlation exceeding the population correlation on a specific trial after the shift, for random samples with sizes 4, 7, and 10.

Conclusion. A strict translation of the small sample hypothesis by using a random sample procedure results with the same qualitative predictions as my decay model. Therefore, I think it is appropriate to map differences in sample size onto differences in decay. At the same time, it appears psychologically more plausible that if people, due to capacity limitations, have to rely on a sample of data, the sample will tend to include more recent cases rather than randomly sampling from all cases. It is simply that older cases are harder to retrieve. This assumption is embedded in ACT-R's mechanisms for retrieval competition and is endorsed by other researchers by its inclusion in their own computational models of cognition (e.g., Erev, 1998; Rieskamp, Busemeyer, & Laine, 2003; Yechiam & Busemeyer, 2005).

1.4.5 Signal detection analyses of the ACT-R models

I was also interested in discussing the ACT-R models in terms of signal detection theory. As Juslin and Olsson (2005) pointed out, it does not suffice to look at the hit rate, but one ultimately needs to consider the posterior probability of a hit, which also takes false alarms into account. More specifically, I was interested in the question whether the

noise and the decay models make different predictions about high and low spans with regard to sensitivity and response bias.

Method. I conducted a simulation of 4 blocks with 32 trials each of the envelope task with two different kinds of envelopes, red and green ones, which contained two different kinds of symbols, Xs and Os. The simulation was run 1000 times with every single model. In the noise condition, there was no correlation. Both symbols had a prevalence of .5 within both the red and green envelopes, and both kinds of envelopes had an equal probability of occurrence. In the signal condition, which was identical to the one described in section 1.4.2, there was a correlation of $\Phi = .375$. One symbol (e.g., X) had a prevalence of .6875 within one kind of envelope (e.g., red) and a prevalence of .3125 in the other (e.g., green), while this was reversed for the other symbol. Again, both kinds of envelopes had an equal probability of occurrence.

Results. Here, I define a model's response bias as its tendency to produce false alarms. That is, to respond to the noise trials as if they were signal trials. The threshold to count a model's responses to a block as a false alarm was a maximizing rate of at least .6875, which corresponds to the prevalence of the more frequent symbols (given a color) in the signal condition.

The false alarm rates are depicted in Figure 1.4. As can be seen, the noise model predicts higher false alarm rates for low spans on all blocks. These simulation results fit nicely with the interpretation that low spans explore less and settle more quickly on maximizing, while the high spans are more careful in drawing conclusions and continue to explore longer. The decay model, in contrast, predicts a higher response bias for high spans, but basically only on the first block, while the difference is very small thereafter.

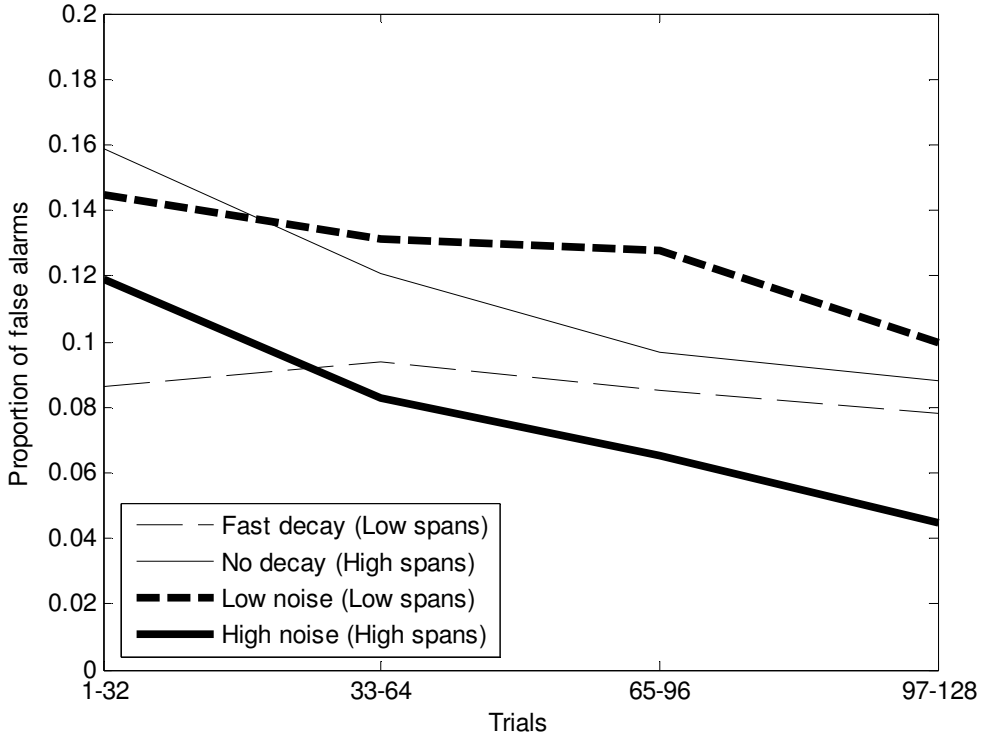

Figure 1.4. False alarm rates on 4 blocks of 32 trials each for the different models.

Differences in sensitivity can be seen in comparing, for each model, how well it can differentiate between signal and noise trials. For the sensitivity analysis, I acquired ROC curves for the 4 blocks by varying the decision threshold that needs to be exceeded before interpreting the models' responses as signal responses. This threshold was varied between 0 and 1 in steps of .1. Figure 1.5 shows those ROC curves, separately for A) the decay models, and B) the noise models. As can be seen, the models representing the low spans (fast decay and low noise) have a higher sensitivity than the models representing the high spans (no decay and high noise).

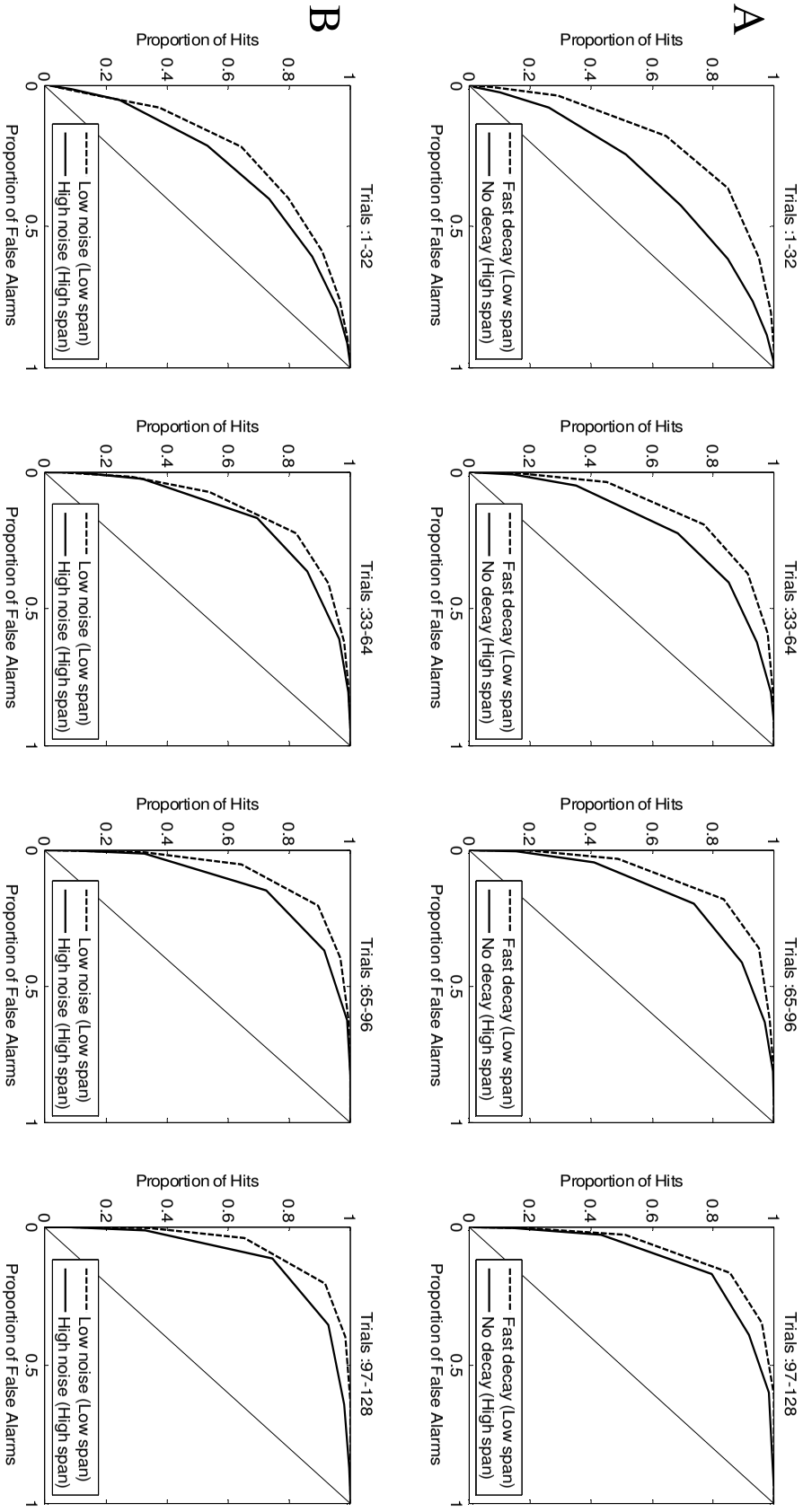

Figure 1.5. ROC curves for the (A) decay models and (B) noise models.

Conclusion. In sum, the signal detection analyses illustrate that both the decay and noise models predict a higher sensitivity for the variant describing low spans. Furthermore, it is interesting that the noise model predicts a higher response bias for low spans, which can be interpreted as less explorative behavior and thus is congruent with the predictive behavior hypothesis.

1.4.6 Overall discussion of the ACT-R models

With the simulations, I tried to make differential predictions between the predictive behavior hypothesis (modeled with noise) and the small sample hypothesis (modeled with decay). Importantly, the decay model is congruent with major ideas of Kareev et al. (1997)'s small sample hypothesis. Even small random samples are more likely to reveal the shift than large random samples, as I found out in supplementary simulations. Thus, also strictly translating the random sample procedure of the small sample hypothesis results in qualitatively identical predictions to my decay model. Therefore, I think it is appropriate to translate the small sample hypothesis with differences in decay *d*.

When fitting the two model variants to the data, the noise variant had a slightly better fit in predicting participants' behavior. I could not improve the fit of the decay variant by only varying decay *d* because the parameter values in the decay variant are at the extremes of the reasonable parameter value space. The extreme values signal a problem with the decay variant because usually they are not set too differently from the default value of $d = .5$. These results already support the predictive behavior hypothesis represented with the noise variant of the model. However, both model variants provide a good fit to Kareev's data. A more decisive comparison can be provided by considering the two qualitatively different predictions that the two variants of the model make after a shift in the environment occurs. Therefore, a correlation detection task (or probability learning task, as I conceptualize it) that includes a shift in the environment will assist in deciding which of the two hypotheses of the low capacity advantage on correlation detection (small sample hypothesis vs. predictive behavior hypothesis) is more likely. If short-term memory capacity affects people's perception of contingencies (or conditional probabilities) in the manner suggested by the small sample hypothesis (Kareev et al., 1997), then it should be captured by the decay parameter. The model makes the clear prediction that this should result in a low capacity advantage after a shift. If, however, lower short-term memory capacity fosters simple maximizing, then the data should be

congruent with the predictions made by varying the noise parameter. Thus, there should be a high capacity advantage after a shift.

Altogether, four Experiments were conducted to test the predictions of the ACT-R model. Experiment 1 and Experiment 2 are the major experiments in testing the predictive behavior hypothesis against the small sample hypothesis. In both Experiment 1 and Experiment 2, people who naturally differ in their short-term memory capacity were compared with respect to their behavior in the correlation detection task used by Kareev et al. (1997). In Experiment 3, memory capacity was experimentally manipulated with a secondary task, which also allows causal conclusions. Furthermore, Experiment 3 used the simple binary choice probability learning situation (as applied by Wolford et al, 2004) to show that the results between a correlation detection and a simple probability learning task are indeed comparable. Also using this setting, Experiment 4 tries to test the result from the signal detection analyses that low and high spans could not only differ with regard to their sensitivity, but also with regard to response bias.

1.5 Experiment 1

Experiment 1 was designed to assess the impact of short-term memory capacity on behavior in an extended version of the correlation detection task used by Kareev et al. (1997, Experiment 1). To test the model predictions empirically, I added shifts in the correlational structure of the task (i.e., reversals of the correlations) which made it necessary to conduct a computer version of the task. To obtain a more complete picture of people's cognitive capacities, I applied measures of working memory in addition to the digit span short-term memory task that Kareev et al. used.

The idea for using these additional working memory measures was that they allow for testing an additional hypothesis, regarding performance after a shift, not captured by the models. Performance after a shift will not only depend on detecting the change but also on how susceptible people are to proactive interference. That is, how strongly information, that they have learnt so far, will interfere when people attempt to learn new information or when they attempt to adapt their behavior to this new information. Kane and Engle (2000) found that people with a low working memory capacity are more susceptible to proactive interference. Therefore, one could imagine that low spans, even if they detected the shift earlier, are not able to adapt their behavior to this shift appropriately because they are more susceptible to proactive interference. Such an effect could negate a possible

advantage, resulting from an earlier detection of correlation. This alternative hypothesis is, in a sense, the opposite of the decay model in ACT-R. While the decay model assumes faster forgetting for low spans, and thereby a recency effect, the proactive interference hypothesis assumes a stronger primacy effect for low spans. That is, it assumes that low spans, due to proactive interference, put too much weight on old information, and thereby fail to adapt to a changing environment.

1.5.1 Methods

Participants. Eighty students (42 female) with an average age of 24 years (SD = 3.5) participated in the experiment. They were paid 7€ for participation, plus a bonus depending on their performance.

Design and procedure. Each participant was tested individually in a quiet room. I retained the task order of the original Kareev et al. (1997) study. First, short-term memory capacity was measured with a digit span forward task (as in Kareev et al.). People were required to verbally repeat sequences of digits that were read to them by the experimenter at a pace of approximately one digit per second. After correct repetition, the length of the sequence increased by one digit, whereas a failure ended the task. Digit span capacity was determined by the highest number of correctly repeated digits. After the digit span forward task, participants were seated in front of a computer, where the correlation detection task was presented to them. This task was a computer adaptation of the correlation detection task used by Kareev et al. (Experiment 1). Participants sequentially encountered red and green envelopes on the computer screen. On each trial, they had to predict whether the envelope contained a coin marked with an X or an O. They received visual feedback lasting 3 seconds after each trial and were paid 3 € cents for each correct prediction. Kareev et al. similarly rewarded their participants. Overall, there were 384 trials divided into three seamless blocks (consisting of 128 trials each), in each of which, envelopes were drawn randomly without replacement[5]. With regard to differences in performance between high and low spans, the conditions with a correlation of $\Phi \approx |.4|$ had, on average, the largest effect size in Experiment 1 by Kareev et al. Therefore, I decided to administer a condition with a correlation of that size. For all participants, the first block in Experiment 1 corresponded to Kareev et al.'s condition with $\Phi = |.375|$ in which the total amount of Xs

[5] I wanted to be as close as possible to Kareev et al.'s (1997) Experiment 1, where people drew envelopes from a real bag, also without replacement.

and Os was equal (i.e., a symmetric condition): Within this block, each participant encountered an identical distribution of color-symbol combinations consisting of 44 Xs (68.75%) and 20 Os (31.25%) in red envelopes, and 20 Xs and 44 Os in green envelopes.

There were four conditions that were identical in the first block, but differed according to whether shifts in the correlational structure (i.e., in the probabilities of outcomes given the color of the envelope) occurred in the second and/or in the third block. A shift always consisted of reversing the correlation, resulting in $\Phi = -.375$. That is, the distribution of symbols within the envelopes was exactly reversed, so that there were 20 Xs and 44 Os in red envelopes, and 44 Xs and 20 Os in green envelopes, in blocks after a shift. This large shift has the methodological advantage of leading to very distinct predictions of the two hypotheses I want to test against each other. Given the probabilistic nature of the task, a more subtle shift than this could have been too difficult for the participants to detect. There was no cue to indicate shifts in the correlational structure.

In the constant condition, no shift occurred, in the early shift condition, a shift occurred after the first block, in the late shift condition, a shift occurred after the second block, and in the back shift condition, there was a shift after the first block and a shift back to the initial correlation after the second block (see Table 1.1).

Table 1.1. Conditions in Experiment 1: Positive or Negative Correlations in the Blocks

Condition	Block 1	Block 2	Block 3
Constant	+	+	+
Early shift	+	–	–
Late shift	+	+	–
Back shift	+	–	+

The motivation of the different conditions was the following. The constant condition is useful to see how the low capacity advantage, if replicable, develops over time. Since I did not know when a change would affect participants strongly, I thought it was useful to also have an early and a late shift condition, independent of which model is more appropriate. If people catch a change in the environment quickly, then it is interesting to see how they detect another change as is provided in the back shift condition.

After the correlation detection task, I administered a counting span and an operation span task (Engle et al., 1999) as additional working memory measures. The main difference between short-term and working memory is that short-term memory only requires storage, while working memory additionally requires processing (Miyake et al., 2001). The counting span task consisted of counting aloud the objects on the screen, and remembering the number for a later test. After several trials, participants had to recall all the numbers from the last two to six trials. For the operation span task people had to evaluate simple mathematical equations, and read aloud words that appeared with the equations on the screen. After two to five trials they had to write down the words from these trials.

1.5.2 Results and discussion

For all analyses, results of different conditions were pooled for blocks that were identical in both position (i.e., first, second, third) and learning history. That is, the analyzed block and all previous blocks had to share the same correlational structure. For example, behavior in block 2 after an early shift can be pooled across the early and the back shift conditions. Table 1.2 summarizes all correlations between the different capacity measures and the relative frequency of maximizing behavior on the different blocks.

Table 1.2. Summary of Results in Experiment 1

	Maximizing						
	Pre-shift			Post-shift			
				early		Late	back
Block	1	2	3	2	3	3	3
Digit span	$r = -.23$, $p = .04$	$r = -.16$, $p = .32$	$r = -.44$, $p = .05$	$r = .13$, $p = .41$	$r = -.04$, $p = .87$	$r = .49$, $p = .03$	$r = -.11$, $p = .64$
Counting span	$r = .01$, $p = .96$	$r = -.15$, $p = .37$	$r = -.21$, $p = .38$	$r = -.14$, $p = .38$	$r = -.08$, $p = .73$	$r = .19$, $p = .43$	$r = -.12$, $p = .61$
Operation span	$r = -.06$, $p = .60$	$r = -.17$, $p = .30$	$r = -.08$, $p = .73$	$r = -.03$, $p = .85$	$r = -.15$, $p = .53$	$r = -.13$, $p = .58$	$r = -.14$, $p = .56$
N	80	40	20	40	20	20	20

Replication. Analyzing the first block, which was comparable for all participants, allowed testing whether the low capacity advantage observed by Kareev et al. (1997) was replicable. In keeping with the original analysis, I split the participants into two groups according to their median digit span capacity. Since it was not clear whether to treat those with median scores as high or low spans, I decided to exclude them. I believe this adds less noise than Kareev et al.'s procedure of randomly categorizing participants with a median value as either high or low digit spans. Low digit spans (M = 73.82, SD = 5.40) performed better on the task than high digit spans (M = 68.75, SD = 9.30), $t(43.37) = 2.50$; $p = .02$, with corrected degrees of freedom due to higher a variance for high spans, $F(1, 54) = 10.68$, $p < .01$. The mean difference corresponds to an effect size of Cohen's $d = 0.67$. This effect size is lower, compared to the corresponding condition of Kareev et al., with an effect of $d = 0.94$. As I deliberately picked a condition with a comparatively large effect size, some regression to the mean is likely to occur. In Kareev et al., the overall effect size was $d = 0.33$. Thus, the effect size in the present study was somewhere between the overall effect size Kareev et al. had observed and that which was observed in the conditions closest to my own. In sum, the original finding could successfully be replicated.

The variance for high digit spans was higher because their prevalence of maximizing was lower, on average. A group of participants that adopted perfect maximizing would have the same expected performance. In contrast, a group of participants that did not adopt maximizing would, on average, perform less well, compared to the maximizing group, but would also show much more variance in performance, which could, in principle, vary between zero and 100% accuracy.

Since the performance depends to a certain degree on chance, I decided to focus on the relative frequency of maximizing. For each participant, I computed the proportion of trials in which participants chose the option corresponding to maximizing (i.e., choosing X if red and O if green before the shift, and vice versa after the shift). A value of .5 reflects random behavior, a value close to the frequency of the more frequent event in the environment (68.75%) reflects probability matching, and a value of 1 reflects perfect maximizing. I argue that this measure is less noisy than the performance because it is independent of the outcome of a decision (although it naturally correlates with performance, $r = .89$, $p < .01$). I think that this measure is easier to grasp intuitively than the measure Kareev et al. (1997) used, which they originally called perceived correlation.

The relative frequency of maximizing is correlated by 1 to perceived correlation, and for my analyses, it made no difference which measure was applied.

In the analysis reported above, I used a median split to correspond with Kareev et al.'s (1997) analysis. However, median splits decrease statistical power and can introduce error, primarily because the inherent variability of the predictor is reduced (Irwin & McClelland, 2003). Therefore, in the following analyses, correlations include all levels of digit span capacity. The low digit span capacity advantage was also reflected in a negative correlation between digit span capacity and pre-shift maximizing on the first block ($r = -.23$; $p = .04$), indicating that low digit spans show maximizing more frequently in this block. The course of pre-shift maximizing on the first 128 trials is depicted in Figure 1.6.

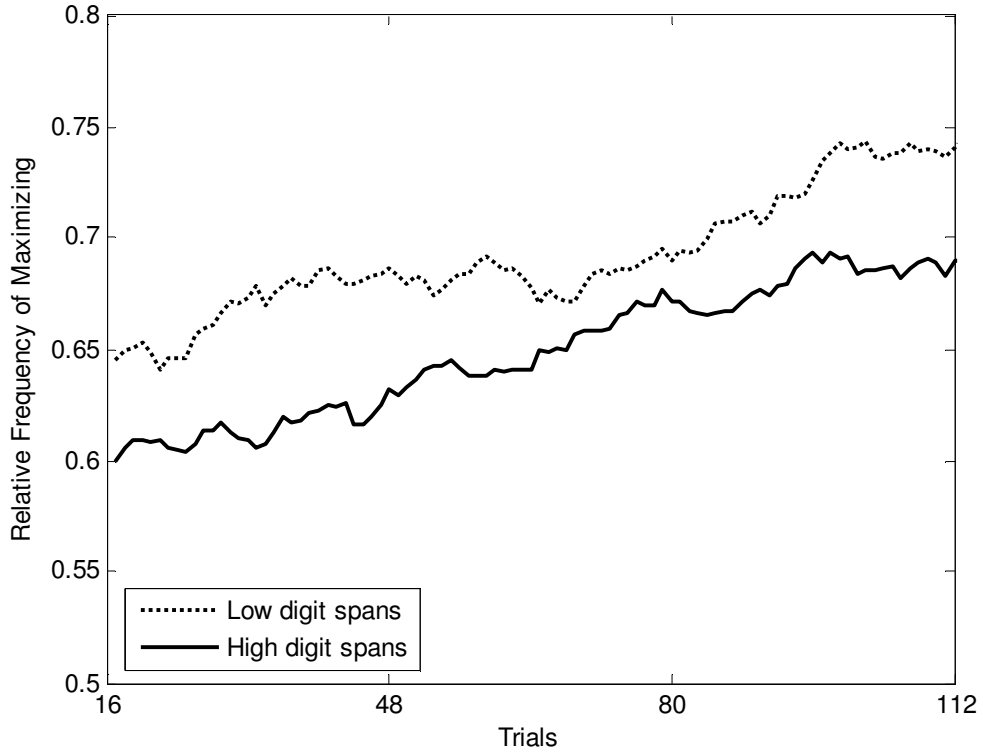

Figure 1.6. Pre-shift maximizing on block 1, Experiment 1. The amount of maximizing is averaged within a moving window of 32 trials and is reported separately for high and low digit spans as derived by the median split.

Post-shift trials. In the trials after a shift, the correlation in the environment was reversed. Therefore, maximizing now consisted of choosing the opposite object, given a color (e.g., O is now the maximizing answer, given red, since the maximizing answer was X previously). In contrast to the small sample hypothesis, there was no relation between

digit span capacity and post-shift maximizing behavior on the early post-shift block (r = .13, p = .41), and even a high digit span capacity advantage, indicated by a positive correlation between digit span capacity and post-shift maximizing on the late post-shift block, was observed (r = .49, p = .03). These results are contrary to the prediction of the decay variant of the model implementing the small sample hypothesis. According to the decay variant model, the low digit span capacity advantage, corresponding to a fast decay parameter value of the model, leads to an even more pronounced advantage after a shift. Instead, the data revealed either no effect or the opposite, and is thereby congruent with the predictions made by the noise variant model representing the predictive behavior hypothesis. Post-shift maximizing behavior was only related to digit span capacity in the late shift condition, and here, the correlation was positive. That is, high digit spans adopted maximizing with a higher relative frequency after the late shift. This condition is depicted in Figure 1.7.

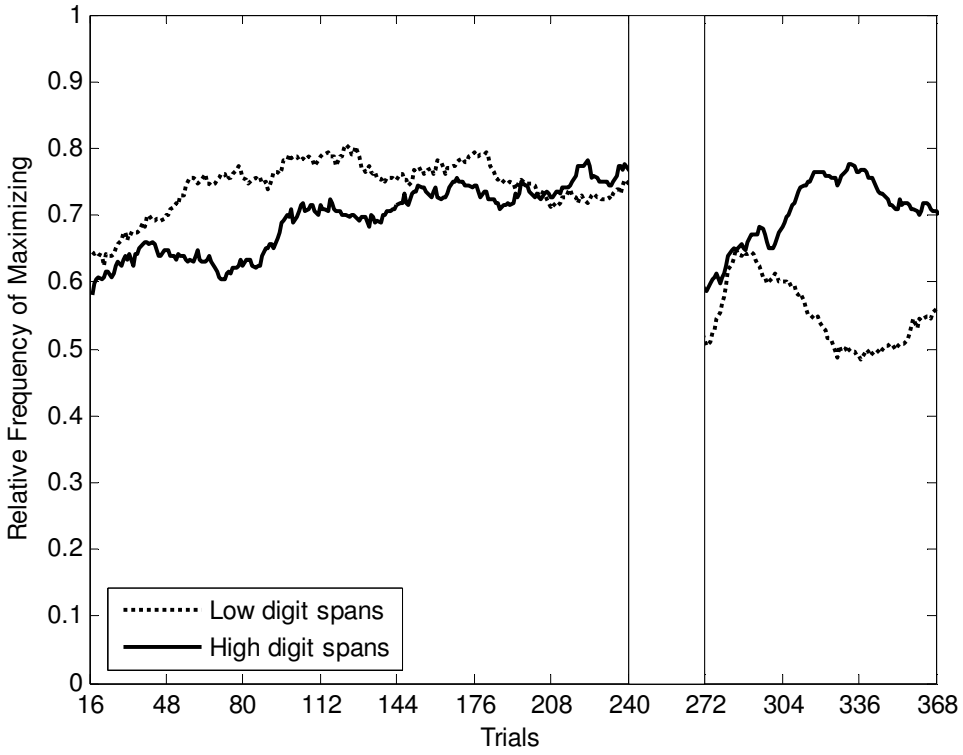

Figure 1.7. Maximizing on all trials, Experiment 1, late shift condition. Low and high digit spans were averaged separately across trials within a moving window of 32 trials. To prevent an overlap between trials before and after the shift in this window, I started averaging again after the shift, which is indicated by the two vertical lines at trials 240 and 272. That is, the last depicted data point before the shift consists of the last 32 trials before the shift, and the first depicted data point after the shift consists of the first 32 trials after the shift.

Other working memory measures. Naturally, digit span capacity was correlated with both counting span (r = .24; p = .03) and operation span (r = .24; p = .03). However, the other working memory measures were unrelated to pre- and post-shift behavior. That is, neither the low digit span capacity on pre-shift trials nor the high digit span capacity advantage on post-shift trials could be captured by those measures. If the high digit span capacity advantage on the post-shift trials were due to higher proactive interference of the low digit spans, this should be captured with one of the other working memory measures, which were also used by Kane and Engle (2000). Therefore, I am confident that this high digit span capacity advantage on post-shift trials indeed favors the predictive behavior hypothesis (although I try to more carefully rule out the proactive interference hypothesis in Experiment 2, see below).

Preliminary conclusion. The overall picture supports the hypothesis that it is not the perception of correlation that differs between people with high and low digit span capacity, but differences in predictive behavior (i.e., differences in how consistently they maximized their payoffs). There was a low digit span capacity advantage before a shift, but no difference or even a high digit span capacity advantage that emerged after a shift. Thus, the data are not at all congruent with the decay variant of the model, but they are congruent with the noise variant, and thereby, my assumption that differences in predictive behavior are of importance.

However, a high digit span capacity advantage after an early shift was not found, but only after a late shift. Since the sample size of the late shift condition in which I found a post-shift high digit span capacity advantage is small (n = 20), this finding should be interpreted with care.

1.6 Experiment 2

The second experiment was a slightly refined version of the first, intended to replicate the important results of Experiment 1. Now, I know that a change in the correlational structure of the environment only reveals differences between high and low spans after many trials. Therefore, I only implemented the late shift condition in which I found a high capacity advantage. I also wanted to more strongly rule out the alternative hypothesis that high spans were at an advantage after a shift because they were less susceptible to proactive interference. In Experiment 1, I only addressed this question by assessing additional working memory measures that were shown to be related to proactive

interference (Kane & Engle, 2000). However, Kane and Engle used extreme group comparisons and a large sample size (192 and 216 participants, respectively) to show the modest relation between working memory and susceptibility to proactive interference. That is, there could have been proactive interference which I did not capture with my working memory measures. Therefore, I assessed susceptibility to proactive interference directly.

1.6.1 Methods

Participants. Eighty students (51 female) with an average age of 24 (SD = 3.6) participated in the study. They were paid 9 € for participation, plus a bonus depending on their performance (identical to Experiment 1, per correct trial 3 € cents).

Design and procedure. Each participant was tested individually in a quiet room. Again, I kept the task order as in the original study by Kareev et al. (1997), starting with the digit span forward task to measure short-term memory capacity. This time, digit strings were digitally recorded beforehand, so that participants listened to identical audio files instead of listening to an experimenter reading the digits to them. The correlation detection task consisted of only the late shift condition of Experiment 1, with a shift seamlessly occurring after two blocks. Colors of the envelopes and keys on the keyboard (e.g., whether X was left or right) were counterbalanced. For a more detailed description of the task, see Experiment 1.

After the correlation detection task, the counting span task (see Experiment 1) was administered. Furthermore, I assessed susceptibility to proactive interference (Kane & Engle, 2000), which I considered to be a possible alternative explanation for the high digit span capacity advantage after a shift in Experiment 1. This task consisted of learning three word lists with words that belong to one category (professions) and one word list that belongs to another category (animal names). The words were presented successively, and participants had to recall as many words as possible after each list. It is usually observed that performance decreases over the course of the three word lists from one category (proactive interference) and then increases again on the last word list (proactive interference release).

1.6.2 Results and discussion

A repeated measure analysis revealed no difference between the counterbalanced conditions with regard to maximizing in the three blocks, $F(5, 126.8) = .76$; $p = .58$.

Therefore, all counterbalancing conditions were merged. Surprisingly, the original low capacity advantage on pre-shift maximizing could not be found in Experiment 2. There was no significant correlation between digit span capacity and pre-shift maximizing on block 1 ($r = –.08$; $p = .50$) and on block 2 ($r = –.10$; $p = .38$). There was also no post-shift high digit span capacity advantage, post-shift maximizing on block 3 was unrelated to digit span capacity ($r = .10$; $p = .39$).

Neither proactive interference nor its release could predict any behavior. That is, post-shift maximizing really does not seem to be a function of susceptibility to proactive interference at all. Proactive interference was not correlated with digit span or counting span. Surprisingly, counting span was positively correlated to pre-shift maximizing on block 2 ($r = .27$, $p = .02$).

Since both experiments were almost identical in structure, this result surprised me. Therefore, I suspected that some peculiarity of my sample in Experiment 2 might be responsible. Digit span capacity and counting span capacity were comparable between the experiments. The only demographic variables assessed were age and sex. The only difference between the samples from the two experiments that struck me was the larger proportion of women in Experiment 2, compared to Experiment 1 (63.8% vs. 52.5%), which suggested that I should explore sex differences in a post hoc analysis.

1.6.3 Post hoc analyses of sex differences

One reason for the different results might be based on sex differences since a different proportion of men and women participated in Experiment 2. I decided to merge the data sets from my two experiments to have a reasonable sample size to analyze men and women separately.

Merging the data sets only makes sense for blocks that are identical in both position (i.e., first, second, third) and learning history for both experiments, which is the case for the first two pre-shift blocks and the late post-shift block. It results in sample sizes of $n = N = 160$ for pre-shift block 1, $n = 120$ for pre-shift block 2, and $n = 100$ for post-shift block 3 from the late shift condition. For all other blocks, I do not have an appropriate sample size to further divide them by sex. Individual difference measures assessed in both experiments were digit span and counting span.

Looking at the correlations between digit span and counting span, on the one hand, and relative frequency of maximizing, on the other hand, separately for men and women, indeed revealed a sex difference. The pre-shift low digit span capacity advantage and the

post-shift high digit span capacity advantage only existed for men, but not for women. For women, there was even a positive correlation between counting span and pre-shift maximizing (see Table 1.3).

Table 1.3. Sex Difference in the Interaction between Capacity and Maximizing

	Maximizing					
	Men			Women		
	Pre-shift		Post-shift late	Pre-shift		Post-shift Late
Block	1	2	3	1	2	3
Digit span	$r = -.19$, $p = .12$	$r = -.43$, $p < .01$	$r = .36$, $p = .03$	$r = -.10$, $p = .37$	$r = .13$, $p = .30$	$r = .12$, $p = .34$
Counting span	$r = .03$, $p = .83$	$r = -.09$, $p = .53$	$r = -.03$, $p = .86$	$r = .18$, $p = .09$	$r = .33$, $p < .01$	$r = .21$, $p = .11$
N	67	50	38	93	70	62

To illustrate this, Figure 1.8 depicts the relative frequency of maximizing, separately for men and women from the late shift condition in Experiment 1 and from Experiment 2 in which only the late shift condition was conducted. Men and women were separately divided into high and low digit spans with a median split (based on all participants), and the relative frequency of maximizing is averaged within a moving window of 32 trials.

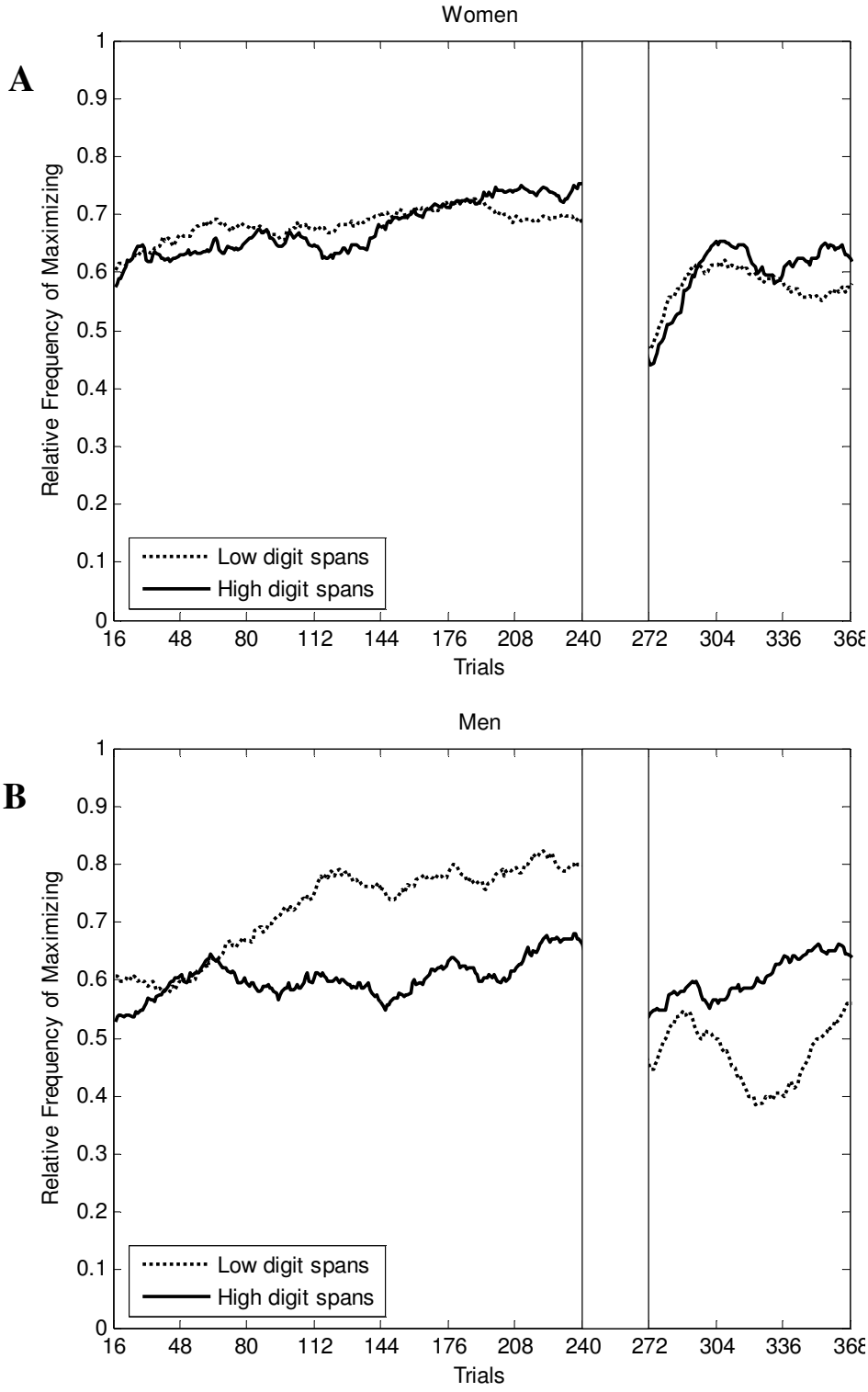

Figure 1.8. Maximizing separately for high and low digit spans, late shift condition, both experiments, for (A) women and (B) men. I averaged low and high digit spans separately across trials within a moving window of 32 trials and started averaging again after the shift, which is indicated by the two vertical lines at trials 240 and 272 (see also Figure 1.7).

The difference lies in the interaction. Men and women did not differ on absolute levels of relative frequency of maximizing (M_{Men} = .64; M_{Women} = .65; F[1, 158] = .22, p = .64) or performance (M_{Men} = 70.82%; M_{Women} = 70.99%; F[1, 158] = .02, p = .89) on block 1, which I chose for this comparison because there are comparable data for all participants on this block. Note, however, that digit span capacity was higher for men (M_{Men} = 6.22; M_{Women} = 5.85; F[1, 158] = 4.32, p = .04) which was not the case for counting span (M_{Men} = .71; M_{Women} = .69; F[1, 158] = .34, p = .56). There was a correlation between digit span and counting span for men and women (r = .26, p = .03, and r = .21, p = .04).

Fortunately, I was able to test whether this sex difference is a peculiarity of my samples or something that may be more general because Kareev provided the original data set from Kareev et al.'s (1997) Experiment 1. It included a total of 112 participants, 64 of whom were women. Note that this experiment did not include a shift in the correlational structure, so that I could only test whether the sex difference on pre-shift trials also holds there. It does: There only is a (negative) correlation between performance and digit span capacity for men (r = –.28, p = .06), but not for women (r = .06, p = .66). The same holds for the correlation between digit span and the absolute strength of perceived correlation (which corresponds to the variable I call maximizing), which only existed for men (r = –.29, p = .05), but not for women (r = -.05, p = .68). Here, men and women did not differ with respect to performance (F[1, 110] = .44, p = .51), the absolute strength of perceived correlation (F[1, 110] = 1.44, p = .23), and digit span capacity (F[1, 110] = 1.54, p = .22).

To summarize, the low digit span capacity advantage on trials before a shift only exists for men, which was the case for the present experiments and Kareev et al.'s (1997) Experiment 1. In Experiment 2, the pre-shift low digit span capacity advantage developed over time and was stronger on pre-shift block 2. In my view, this strengthens my argument that the difference between high and low digit spans lies in differences in predictive behavior. If the difference lied in perception, and thereby in the earlier detection of the correlation by low digit spans, as assumed by Kareev et al., then this difference should be more pronounced earlier, rather than later. The post-shift high digit span capacity advantage also existed only for men. That is, for women, digit span was unrelated to behavior. But interestingly, counting span was related to behavior, but in the opposite direction: It was positively correlated to the relative frequency of maximizing before a shift.

1.7 Discussion of Experiments 1 and 2

Although there are two more experiments to come in this chapter, I think it is worthwhile to discuss the major findings and the major peculiarities of Experiments 1 and 2 already here because the following experiments, although based on the predictions of the same model, will be different in design and procedure.

The goal of Experiments 1 and 2 was to disentangle two potential explanations for the stunning finding of a low digit span capacity advantage on correlation detection (Kareev et al., 1997). Kareev et al.'s original explanation was that low digit spans perceive correlations as more extreme than they actually are because they based their estimates on smaller samples from the environment. Small samples statistically tend to overestimate correlations, and this overestimation can be advantageous in correlation detection. I have called this the small sample hypothesis.

However, the small sample hypothesis has been criticized theoretically (R. B. Anderson et al., 2005; Juslin & Olsson, 2005), and some studies dealing with contingency assessment provide conflicting evidence (e.g., Clément et al., 2002; Shanks, 1985, 1987). Therefore, I explored whether the low digit span capacity advantage found by Kareev et al. (1997) could be explained differently. Instead of assuming that people differ in their perception of correlations, I assumed that people differ in their predictive behavior. This predictive behavior hypothesis was inspired by revisiting the related probability learning literature, which revealed convergent evidence showing that the most successful predictive behavior (maximizing) can be related to a reduced or limited memory capacity.

I therefore hypothesized that low digit spans are more likely to maximize their payoffs more consistently, resulting in the low digit span capacity advantage. Based on the initial learning trials as implemented by Kareev et al. (1997), one cannot distinguish conclusively between the small sample hypothesis and the predictive behavior hypothesis. However, by instantiating both of these explanations in ACT-R models, it was possible to demonstrate that these hypotheses make different predictions about how participants will behave after a shift in the correlational structure of the environment. The model that implemented the small sample hypothesis predicted a low digit span capacity advantage before and after a shift. In contrast, the model that implemented the predictive behavior hypothesis exhibited a low digit span capacity advantage before a shift, but a high digit span capacity advantage after a shift.

1.7.1 Support for differences in predictive behavior

The results of Experiments 1 and 2 replicate the low digit span capacity advantage found by Kareev et al. (1997), although these and the following results only held for men (see section 1.7.4 "A Puzzling Sex Difference"). After a shift in the environment, however, I found either no difference between high and low digit spans or a high digit span capacity advantage. This is contradictory to the assumption made by the proponents of the small sample hypothesis that high and low digit spans differ in their perception, but it is congruent with my assumption that this difference lies in predictive behavior.

However, if one wanted to keep the perceptual argument, one could argue that low spans are less likely to engage in further sampling, because they perceive the correlation as more extreme, and hence reach a conclusion faster. High spans, in contrast, keep sampling, because they have observed a weaker correlation and are less committed to their estimate, and hence are less at a disadvantage when a change takes place. Nevertheless, I see an advantage of my explanation is its consistency with similar findings of a low capacity advantage in the binary choice probability learning literature (e.g., Wolford et al., 2004). The typical binary choice task is very similar to the task at hand, but a sample based perceptual argument cannot hold, because one only needs to acquire information about proportions. In contrast to correlations, sample proportions are unbiased estimators of population proportions. Thus, my account covers those highly related results, which the perceptual argument does not. Moreover, the results for men revealed that the low capacity advantage developed over time (at least in Experiment 2). It was stronger on block 2 than on block 1, which is, in my view, further counterevidence for the small sample hypothesis. According to it, the low digit span capacity advantage lies in the early detection of strong correlations due to the small sample bias to overestimate these correlations. Thus, it should plausibly have the largest effect early in the experiment.

1.7.2 Estimation versus prediction

I think that explaining the low capacity advantage in this correlation detection task with differences in predictive behavior rather than with differences in the perception of correlation also reconciles this low capacity advantage with apparently conflicting empirical evidence. the assumption that people with a lower short-term memory capacity consider smaller samples, and thereby perceive correlations as more extreme, conflicts with findings that correlation estimates are higher with larger rather than with smaller

samples (e.g., Clément et al., 2002; Shanks, 1985, 1987). For correlation estimation tasks, there is usually also no low capacity advantage reported, but rather the opposite (e.g., Shaklee & Mims, 1982). But estimation and prediction are two different processes. For the task at hand, precise estimation is not necessary. It suffices to figure out which symbol is more frequently associated with which color. Building on that, the simpler predictive behavior by low spans is more successful, at least until the shift.

1.7.3 Why is digit span a better predictor than other working memory measures?

I was surprised to only find a relation between behavior and short-term memory capacity assessed with a digit span test, but not with one of the working memory measures, counting span and operation span. It is unlikely that this is due to a lack of reliability in these measures, since these tasks usually have a reliability, based on internal consistency, between .70 and .90 (with 0 and 1 being the borders "no reliability" and "perfect reliability"; Conway et al., 2005). As pointed out before, short-term memory tasks tend to emphasize the simple storage of information, whereas working memory tasks require the additional processing of the information being stored (Miyake et al., 2001).

I have argued that the lower performance of high digit spans is the result of their more complex predictive behavior. They search for patterns, but since there are none, they fail. Of my capacity measures, only digit span capacity was related to the prevalence of maximizing, suggesting that simple storage is particularly important for pattern search. If one adopts the strategy of rehearsing the sequence of events to search for patterns in it, then one constraint on pattern search is storing the sequence. Other processes involved in pattern search, such as hypothesizing about specific patterns, should be more strongly related to working memory (for the relation between working memory and hypothesis generation, see, e.g., Dougherty & Hunter, 2003). Finding no relation between working memory and maximizing, however, suggests that this part of pattern search, at least for this task, did not tax the working memory capacity of even my low span participants. So, the difference between simple storage (required for the digit span test) and more complex processing (additionally required for the working memory tasks) would explain why I only find a relation between digit span capacity and behavior.

1.7.4 A puzzling sex difference

I found an intriguing sex difference in the interaction between digit span capacity and predictive behavior. Only men exhibited the low digit span capacity advantage before

a shift and the high digit span advantage after a shift. In contrast, short-term memory capacity did not explain any variance in the behavior of women. This sex difference in the interaction between short-term memory and predictive behavior exists in the data from Experiments 1 and 2 and in Kareev et al.'s (1997) data.

In the probability learning literature, a sex difference with respect to the absolute amount of maximizing has been reported: West and Stanovich (2003) found that men were more likely to deliberately opt for a maximizing strategy when a typical probability learning task was described to them and they had to specify, in advance, what they would do. Furthermore, there are reports of sex differences favoring men in a similar task, the Iowa Gambling Task (Overman, 2004; Reavis & Overman, 2001). However, in the data reported here, there is no sex difference in decision making (here: maximizing behavior) per se, but only in the interaction between short-term memory capacity and maximizing.

Since maximizing behavior is comparable on average, it is likely that men and women do not differ with regard to the complexity of their predictive behavior on average. However, looking at the low digit spans only, it seems to be the case that women are better able to engage in complex behavior (resulting in non-maximizing) despite a low digit span capacity. This suggests that women can draw on resources other than simple storage capacity, while men draw more exclusively on the simple storage that the digit span test presumably taps.

Since I do not have additional data, I can only speculate about what these resources could be. Reliable sex differences favoring women have been repeatedly shown on episodic memory tasks, particularly those with a verbal component (for an overview, see Herlitz, Nilsson, & Bäckman, 1997). More generally, females surpass males on tasks in which verbal processing of material is either required or can be used (Lewin, Wolgers, & Herlitz, 2001). This indicates that women could more easily engage in verbal processing to solve a task. Speculating about patterns in sequences of events is a task where verbalization clearly is possible. Thus, women could draw on verbal episodic memory to search for patterns, while men are apparently more likely to depend on short-term memory to store these sequences. This would explain why digit span capacity only explains the variance in the maximizing behavior for men, but not for women.

1.7.5 Limitations

One potential drawback of both Experiment 1 and Experiment 2, however, is that they were only of quasiexperimental nature because they related natural differences in

short-term memory capacity to behavior. Hence, it is, in a strict sense, not appropriate to assume a causal connection between the limitations in short-term memory capacity and the predictive behavior. Therefore, I decided to administer a simple binary choice probability learning experiment with a secondary verbal working memory task, building on the design of Wolford et al. (2004), and to include a change in the environment (i.e., a reversal of the probabilities). This is on the one hand an experimental manipulation that allows drawing causal conclusions. Furthermore, it is a further test of the assumption that the results on the correlation detection task as applied in Experiment 1 and Experiment 2 are indeed comparable to results from binary choice probability learning tests.

I hypothesize that the secondary verbal working memory task which reduces pattern search and thereby increases maximizing, as shown by Wolford et al. (2004), will put people at a disadvantage if the environment changes. In Experiment 1 and Experiment 2, short-term memory capacity could only explain variance in the behavior of men, but not in the behavior of women. The same held for the data of Kareev et al. (1997). Therefore, I was interested whether the secondary verbal working memory task affects men, but not women. Since it thus is clear that I will have to split the sample into men and women and analyze those subsamples separately, I made sure that there were at least 30 men and 30 women in each of the conditions (cognitive load vs. no load).

1.8 Experiment 3

1.8.1 Methods

Participants. 122 students (62 female) with an average age of 25.63 years (SD = 3.21) participated in the study. They were paid 10 € for participation plus a bonus depending on their performance.

Design and procedure. Each participant was tested individually. The main task for all participants was a repeated binary choice task, which was an extended version of the task used by Wolford et al. (2004). People had to predict whether a square appeared on the upper or the lower half of the screen for a total of 600 trials. For easier discrimination of the two squares, the upper squares were red while the lower squares were green. On the first 350 trials, the probability of occurrence of an upper red square was p = .75 while it was only p = .25 for a lower green square. On the remaining 250 trials, these probabilities were reversed, that is, p = .25 for an upper red square and p = .75 for a lower green square.

I will refer to this reversal of probabilities as shift in the following. The number of trials was chosen for the following reasons. In Experiment 1 by Wolford et al., a difference between two conditions (see below) reliably emerged after 300 trials. Therefore, I wanted to administer at least 300 trials before the shift. On the other hand, I did not want the participants to be completely exhausted when the shift occurred to ensure that they have a chance to capture it.

People were randomly assigned to one of two conditions. The 3-Back condition consisted of the binary choice task with a secondary verbal working memory task as used by Wolford et al. (2004). This secondary task was a 3-back task. Each time the participants had to predict the square, they saw a digit between 0 and 9 on the screen. On random trials, participants were probed and had to recall the last three digits. They were probed approximately 5 times out of every 100 trials. These trials were randomly selected, with a minimum of three trials between the probes. In addition, the very last trial was probed. The control condition consisted only of the binary choice task without the secondary task.

Participants earned 1 € cent for each correct trial. Consistent with Wolford et al. (2004), this payment was reduced by 20 € cents for each incorrect probe on the 3-Back task.

1.8.2 Results

I wanted to check whether participants in the 3-Back condition paid attention to the secondary task. The mean accuracies (with standard deviations in parentheses) from block 1 to 6 were .89 (.16), .92 (.13), .96 (.09), .95 (.11), .96 (.09), .96 (.08). That is, performance on the secondary task was very good and even slightly higher than in Wolford et al. (2004) where the mean accuracy on the secondary task was .85 (.07).

I compared the average relative frequency of maximizing between the 3-Back and the control condition over the blocks with a repeated measure ANOVA, separately for the trials before and after the shift. Before the shift, there was no between subjects effect for condition ($F(1, 122) = 0.19$, $p = .67$). Nor was there a linear contrast effect for the interaction between block and condition ($F(1, 122) = 1.45$, $p = .23$), which was reported by Wolford et al. (2004) as reflecting a growing separation between the conditions. After the shift, the mean relative frequency of maximizing was lower in the 3-Back condition compared to the control condition ($F(1, 122) = 22.28$, $p < .01$). That is, the secondary task had a strong impact on mean maximizing after the shift in the predicted direction. But

before the shift, at least on the mean level, I could not replicate the findings by Wolford et al. (2004) that the secondary task increases maximizing.

Are there more maximizers in the 3-Back condition? However, a closer look at the data reveals that the means hide what is really going on. When looking beyond the mean level at the relative frequency of participants who maximize consistently, then it becomes apparent that there are many more consistent maximizers in the 3-Back condition. Consistent maximizing is defined, following Wolford et al. (2004), as choosing the maximizing answer on 95% or more of the trials within a block. Figure 1.9 depicts mean maximizing in the upper part and the relative frequency of participants who consistently maximized in the lower part, separately for the two conditions. The vertical line indicates the shift. As can be seen, the mean values do not differ before the shift, but the relative frequency of maximizers is quite different. On the last two blocks before the shift, there are 31/62 maximizers (i.e., people choosing the maximizing answer on 95 or more of these 100 trials) in the 3-Back condition, but only 16/60 maximizers in the control condition ($\chi^2(1, 122) = 7.01$, $p < .01$).

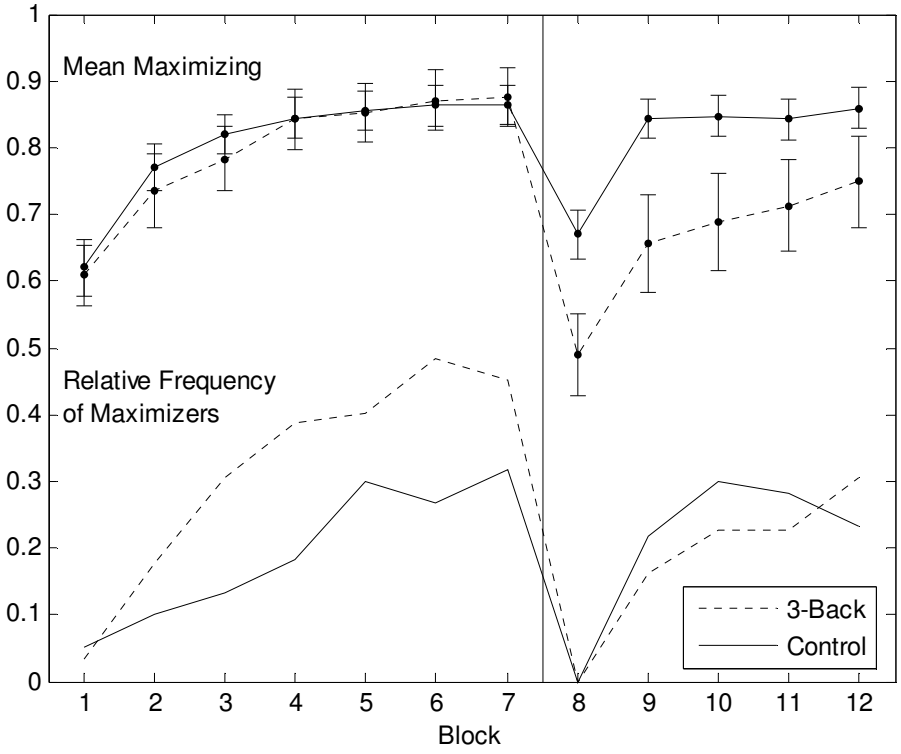

Figure 1.9. Mean relative frequency of maximizing (+ 2 standard errors of mean; upper half) and relative frequency of participants who chose the maximizing answer 95% or more trials of the respective blocks (lower half), both depicted separately for the 3-Back and the control condition.

Another way analyzing maximizing behavior was reported by Shanks, Tunney, and McCarthy (2002). They counted the number of people who had streaks of choosing the maximizing answer on more than 50 consecutive trials. Before the shift, there were 35/62 participants who had such streaks in the 3-Back condition, but only 19/60 in the control condition ($\chi^2(1, 122) = 7.59$, $p < .01$). The mean length of streaks before the shift was M = 99.29 (SD = 91.04) in the 3-Back condition and M = 64.67 (SD = 78.47) in the control condition ($t(120) = 2.25$, $p = .03$). After the shift, there were no such differences, neither for the relative frequency of people who maximized consistently (19/62 in the 3-Back condition vs. 25/60 in the control condition, $\chi^2(1, 122) = 1.61$, $p = .21$) nor for the mean length of streaks (M = 54.40 (SD = 61.70) in the 3-Back condition versus M = 59.67 (SD = 57.00) in the control condition, $t(120) = -.49$, $p = .63$).

Taking these results together – more people who maximize consistently before the shift in the 3-Back condition, but no difference in mean maximizing – indicates that there must be also more people with very low mean maximizing in the 3-Back condition. Pooling the mean maximizing across all blocks, separately for blocks before and after the shift, reveals that the distribution of behaviors is much more skewed in the 3-Back condition. On the pooled behavior before the shift, the standard deviation in the 3-Back condition is SD = .15, while it is only SD = .09 in the control condition ($F = 8.13$, $p < .01$). On the pooled behavior after the shift, this difference is even more pronounced: the standard deviation in the 3-Back condition is SD = .24, while it is only SD = .09 in the control condition ($F = 37.26$, $p < .01$). In the 3-Back condition, some people obviously did not react at all to the shift. In the control condition, however, there are quite some people who reach probability matching performance again immediately after the shift. Figure 1.10 depicts histograms of the relative frequency of maximizing across all blocks, separately for the conditions, and thereby illustrates this skewness. The means and standard deviations on the single blocks are also reported.

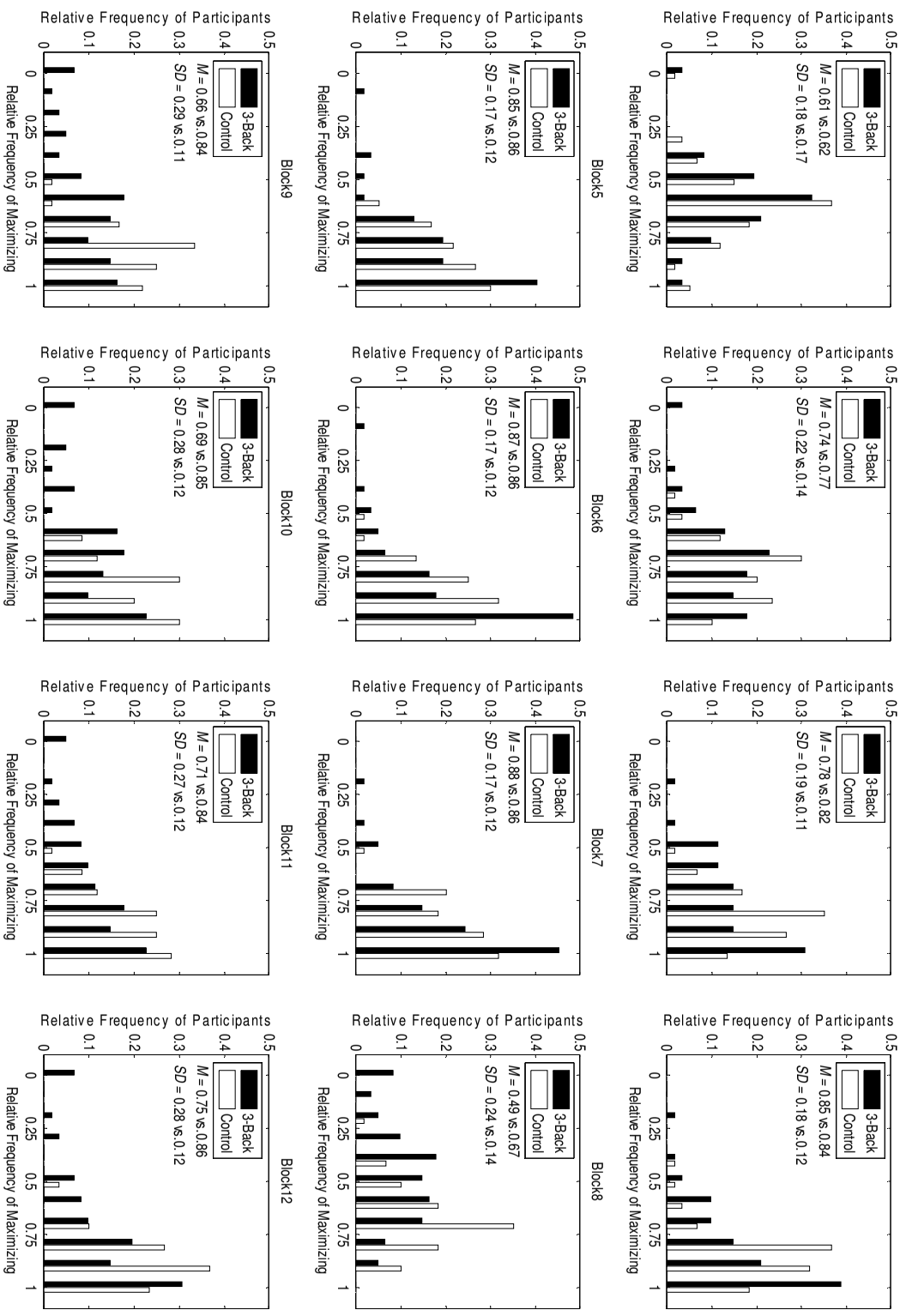

Figure 1.10. Histograms for the relative frequency of maximizing in all blocks, depicted separately for the 3-Back and the control condition, as well as the means and standard deviations. Note that the shift in the environment occurred after block 7.

Sex Differences. Applying many different methods, I did not find any sex differences in this experiment. Neither including gender as a factor in an ANOVA where possible resulted in a significant effect of this factor or in a significant interaction between sex and another factor. Nor did running the chi-square tests for the number of people who maximize consistently on the last 100 trials before the shift or who have streaks of 50 or more consecutive trials choosing the maximizing answer separately for both sexes reveal any differences. Some of those tests fell below standard levels of significance due to smaller sample size, though. But since they are all pointing into the same direction for both sexes, I conclude that there are no sex differences.

1.8.3 Discussion

The results replicate the finding reported by Wolford et al. (2004) that a secondary verbal working memory (3-Back) task increases the likelihood that a person will maximize. I did not find this effect on the level of mean maximizing, though. However, in the 3-Back group there are substantially more people who maximize consistently, but there are also people who pick the less frequent option in the binary choice task more often than the more frequent one and thereby pull the mean down. Thus, the distribution of behaviors in the 3-Back group is much more skewed than in the control group.

Furthermore, the pattern of results is very similar to the results of Experiment 1 and Experiment 2. Taxing memory capacity with a secondary task results in stronger maximizing behavior before a shift, but puts people at a disadvantage after the shift. Thus, I conclude that taxing memory capacity causally affect predictive behavior in the sense that they push people towards more simple and deterministic maximizing behavior instead of searching for patterns. Moreover, the similarity between the results of Experiment 3 to both Experiments 1 and 2 supports the claim that the correlation detection task used by Kareev et al. (1997) can be indeed interpreted as a probability learning task.

There was no sex difference in this experiment regarding the impact of the 3-Back task on behavior, in contrast to the sex difference in the interaction between short-term memory capacity and mean maximizing. I believe that the 3-Back task prevents pattern search much more strongly than a low short-term memory capacity and therefore the effect is observable for both sexes. Note, however, that it also could be that the short-term memory task, which was a digit span task, is solved differently by men and women and thereby measures different things for men and women, while the 3-Back task affects them in the same way.

1.9 Experiment 4

In Experiments 1 to 3, I have studied behavior in environments in which there was a correlation in the data. Although there were no systematic patterns in the sequence of events, participants could learn that there were events that were more prevalent than others. I could demonstrate that in those environments – at least before the environment changed –, a lower or reduced short-term memory capacity was beneficial because it prevented participants from searching for patterns where there are none and instead allowed them to quickly settle on maximizing. In other words, I have shown that participants with a lower or reduced short-term memory capacity have a higher hit rate. That is, they are more likely to jump on something given that there is something to jump on.

Given that the predictive behavior hypothesis (modeled by noise) is indeed right – and all the data so far point in this direction – then low spans should also be more prone to the risk of false alarms. At least the signal detection analyses above (see section 1.4.5) have shown that a lower or reduced short-term memory capacity results not only in a higher hit rate, but also in a stronger response bias, resulting in a higher risk of false alarms. That is, low spans (or people distracted by a secondary task) should also be more likely to jump on something given that there is actually nothing. Experiment 4 is designed to test this prediction in a binary choice probability learning task in which the two events were equally prevalent. Following the signal detection analyses, the hypothesis is that people will even search for patterns in such a task in which there is absolutely no signal as long as they are not distracted by a secondary task. A secondary task distracting people should, similar to Experiment 3, result in more deterministic behavior, indicative of an absence of pattern search behavior.

1.9.1 Methods

Participants. 80 people (40 female) with an average age of 24.71 years (SD = 4.32) participated in the study, most of them were students. They were paid 5 € for participation plus a bonus depending on their performance.

Design and procedure. Each participant was tested in a quiet room. Such as in Experiment 3, the main task for all participants was a repeated binary choice task, which was similar to the task used by Wolford et al. (2004). This time, people had to predict on overall 400 trials whether a square appeared on the upper or the lower half of the screen. For easier discrimination of the squares, the upper squares were red while the lower

squares were green. Both events had an equal prevalence of 50%. That is, there is absolutely no signal one could exploit in any way; any strategy will perform equally well (or badly) on average.

People were randomly assigned to either a 3-back or a control condition, such as in Experiment 3. The 3-Back condition consisted of the binary choice task with a secondary verbal working memory task as used by Wolford et al. (2004). This secondary task was a 3-back task. Each time the participants had to predict the square, they saw a digit between 0 and 9 on the screen. On random trials, participants were probed and had to recall the last three digits. They were probed on randomly selected trials, approximately 5 times every 100 trials, with a minimum of three trials between the probes. The control condition consisted only of the binary choice task without the secondary task.

In addition to the 5€ show-up fee, participants earned 1 € cent for each correct trial. Consistent with Wolford et al. (2004), this payment was reduced by 20 € cents for each incorrect probe on the 3-Back task.

1.9.2 Results

Again, I first wanted to check whether participants in the 3-back conditions paid attention to the secondary verbal working memory task. This was the case: the average accuracy on the secondary task in the 3-back condition was again very high (M = 87.9%, SD = 13.9%).

Since in this experiment both events had an equal prevalence of 50%, there was no maximizing answer. I believe that the behavior of participants who look for patterns should be close to choosing each option in 50% of the cases. Should a person either realize that the sequence of events is random and that both options are equally prevalent or should a person give up the search for patterns, this person's behavior should be more deterministic; and I think that more deterministic behavior will be, on average, closer to choosing one option 100% of the cases. The clear prediction is that the behavior in the 3-back group should be more deterministic on average because participants in that condition should be more likely to give up the search for patterns.

Thus, I simply computed, for each person, how often this person decided for each option. As an indicator of behavior of each person's behavior, I then took the choice prevalence for the option this person has decided for more often. Therefore, this value could vary between 50% if the person has decided for both options equally often and 100% if the person always chose one option, no matter which of the two options this was.

This indicator was computed separately for the first and the second half of the experiment (i.e., for trials 1 to 200 and for trials 201 to 400).

Comparing how deterministic the behavior was in the different conditions, separately for the first and the second half of the experiment, revealed that there was basically no (or only a weak) difference in the first half (3-back: M = .60; SE = .018; Control: M = .57, SE = .016%; $t(78)$ = -1.28, p = .202). On the second half, however, behavior in the 3-back condition was indeed more deterministic than in the control condition, as predicted (3-back: M = .64; SE = .021; Control: M = .58, SE = .017; $t(78)$ = -2.11, p = .038). These results, separately for both halves of the experiment, are depicted in Figure 1.11. There is a rather substantial effect size in the second half of the experiment, d = 0.47.

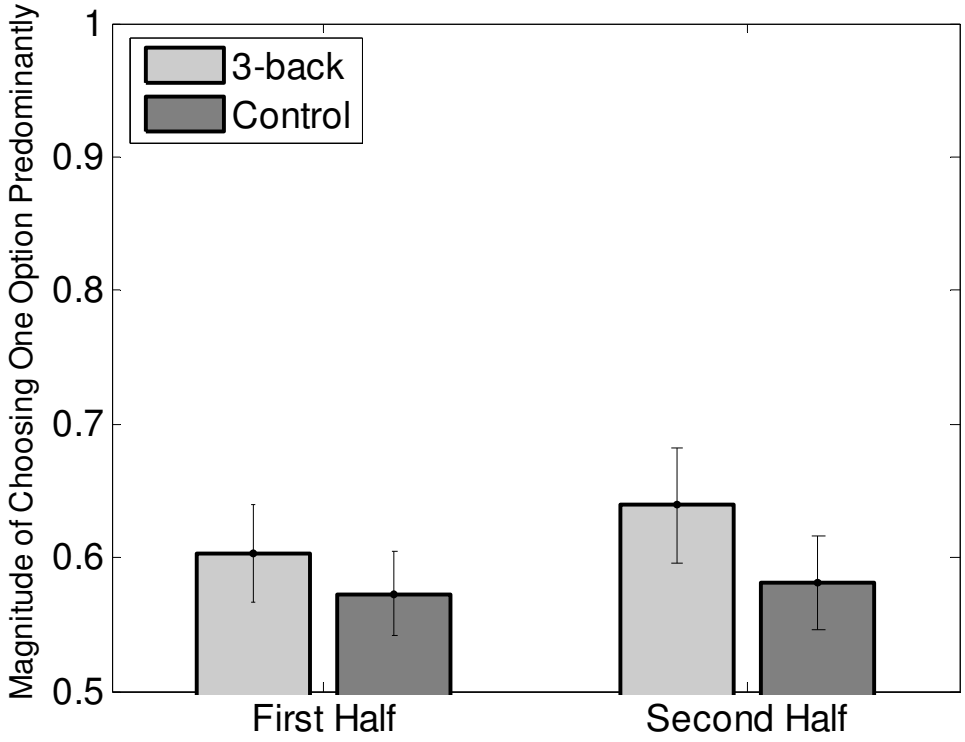

Figure 1.11. Observed magnitude of choosing one option predominantly, separately for both conditions and for both halves of the experiment. The errorbars represent two standard errors of the mean. Since only the magnitude of the predominantly chosen option is plotted, no matter which option this was, a value of 0.5 represents the minimum possible value.

1.9.3 Discussion

Experiment 4 demonstrated that the magnitude of response bias is indeed higher if memory capacities are reduced by a secondary task, such as predicted by the noise model. Response bias was defined here as choosing one option predominantly. This could be interpreted as a kind of false alarm: Participants with lower or reduced memory capacities have a response bias, a tendency to jump on something although there is nothing. In contrast, participants with a higher (or non-reduced) memory capacity match the marginal probabilities more closely, which I interpret as a stronger exploration of potential patterns in order to outsmart the task.

1.10 Overall Discussion

Experiments 3 and 4 provided additional support for the predictive behavior hypothesis, beyond the support that already Experiments 1 and 2 provided. Experiment 3, by experimentally manipulating cognitive load, demonstrated that taxing memory capacity with a secondary task is comparable to naturally occurring differences in short-term memory capacity. Maximizing fueled by the secondary task comes with the price that an adaptation to a changing environment is prevented (or at least slowed down strongly) also in a repeated binary choice probability learning setting. Thus, the claim that the correlation detection task as used by Kareev et al. (1997) can be reinterpreted as probability learning task is substantiated. This also shows that the results by Kareev et al. are indeed comparable to results from the probability learning literature as reviewed above. Furthermore, Experiment 3 demonstrates that taxing memory capacities with a secondary task does not only fuel simple predictive behavior for men (as was the result in Experiments 1 and 2, and in Kareev et al.), but also for women. I have speculated that the sex differences in Experiments 1 and 2 resulted from women being able to draw on different resources (such as verbal episodic memory) to search for patterns, even if their digit span capacity is low. Manipulating cognitive load with a secondary task is the much stronger manipulation compared to naturally occurring differences in digit span capacity. I believe that this strong manipulation also prevents women to draw on other resources to search for patterns, which makes the sex difference disappear.

Experiment 4 provided support for the prediction that a lower or reduced memory capacity should not only yield a higher hit rate (i.e., more successful behavior given that

there is a signal in the environment), but also results in a stronger response bias, including a higher risk of false alarms. In a binary choice situation with both events being equally likely (i.e., 50% each), participants who were distracted by a secondary task behaved more deterministically (i.e., were more likely to predominantly decide for one option) than undistracted participants whose behavior matched the marginal probability of 50% of both events more closely. I interpret the more deterministic behavior of people who are distracted as less explorative – they give up the search for patterns and switch to simpler predictions. The less deterministic behavior of people who are not distracted, in contrast, could be interpreted as them continuously trying to improve their predictions by trying to find patterns in the sequence of events. That is, they are more careful in drawing conclusions and continue to explore longer.

To summarize, the major predictions by the predictive behavior hypothesis, modeled with noise in ACT-R, have been confirmed. First, lower or reduced capacities (modeled with lower noise) result in more successful maximizing behavior as long as the environment is stable. Second, as soon as the environment changes, lower or reduced capacities put people at a disadvantage. These results both hold for naturally occurring differences in digit span capacity and for manipulation cognitive load with a secondary task. Furthermore, these results both hold for the correlation detection task as used by Kareev et al. (1997), which I regard as a probability learning task with two cues and events, and for a simple binary choice probability learning task. Last, Experiment 4 supported the model's additional prediction that lower or reduced capacities should result in a stronger response bias (i.e., being more prone to give up searching and jumping on something although there is nothing). Overall, I consider the noise model representing the predictive behavior hypothesis to be successful. In the next section, I want to discuss why I believe that such a model is indeed plausible.

1.10.1 Plausibility of the ACT-R model

The results I found are congruent with the noise variant of the ACT-R model which I use to implement the predictive behavior hypothesis. Noise has been used in ACT-R to account for different levels of explorative behavior (Taatgen et al., 2006). The noise parameter thereby can account for differences in predictive behavior, capturing the result that the behavior of high digit spans tends to match the probabilities whilst low digit spans are more likely to maximize.

At first glance, it may not be clear how the noise parameter relates short-term memory capacity to the tendency to match the probabilities. The connection may be that people do not actually attempt to match probabilities, but rather that probability matching is the result of more complex predictive behavior, such as pattern search (e.g., Wolford et al., 2004). Looking for patterns in the envelope task and in the binary choice task requires tracking the order of events, which could place high demands on memory. The memory demands of searching for patterns could be even higher than the memory demands for assessing correlations (in the envelope task). Explicitly assessing a correlation requires just knowing the frequencies of color-symbol combinations, since the order in which these appeared is irrelevant. So the cause of the low span advantage could be that people with a low short-term memory capacity find it difficult to entertain very complex patterns and, therefore, tend to settle on maximizing. Since noise is used to model the outcome of either simple predictive behavior (maximizing) or of rather complex predictive behavior (pattern search, resulting in probability matching), I think it is reasonable to capture this short-term memory phenomenon with noise in ACT-R.

It is plausible that more complex predictive behavior helps people to adapt to a shift appropriately. Here, I modeled what I interpret as rather systematic exploration with higher levels of noise. However, there could also be unsystematic, random variability, which can be detrimental. This could be captured by too high levels of noise. Then the predicted behavior would be approximately at chance level before and after a shift. Results that could be interpreted in this way were reported by Sanford and Maule (1973). They compared the relative frequency of maximizing between young and old people in a simple binary probability learning task, including a shift. Older people performed worse than young people before and after the shift. While young people reached probability matching or slightly overmatched both before and after the shift, old people stayed below the probability matching level even before the shift and were approximately at chance level after the shift. This could indicate too high levels of noise (in the sense of too much random variation) for the old people, which clearly slows the adaptation to a shift.

1.10.2 Relations to other models

The implementation in ACT-R is based on the idea that instances of possible solutions are stored in memory that can be used to solve future trials (Logan, 1988). Therefore, it is interesting to consider similarities and differences to other instance based models such as exemplar models. One exemplar model that has been used to explain

multiple-cue probability judgment is ProbEx (probabilities from exemplars, Juslin & Persson, 2002). ProbEx could plausibly solve the envelope task by storing the outcome of each trial as a separate exemplar, with the exemplars containing information about the color of the envelope and the symbol within it. On each trial, it activates exemplars as a function of their similarity to the stimulus. Since each stimulus has only one feature (the color of the envelope), there are always a large number of exemplars that have exactly the same similarity. Thus, it would basically retrieve exemplars proportionally to their frequency of occurrence. ProbEx has a deterministic choice rule and will always select the option supported by more exemplars. But since the sampling of exemplars is probabilistic, ProbEx still could predict behavior similar to probability matching on the aggregated level. Another possibility for ProbEx to approach the task would be to store not only the outcomes as exemplars, but to store the answers as additional exemplars. This would make it similar to my ACT-R models, in which the activation of a chunk is also a function of outcome and behavior. In this manner, ProbEx would predict overmatching and asymptotically approach maximizing over time, because the proportion of exemplars representing the maximizing answer will grow steadily the more often it is chosen. Similar to my model without decay, ProbEx does not forget exemplars and so after the shift would have difficulty overcoming its initial response tendencies.

The exemplar based random walk model (EBRW, Nosofsky & Palmeri, 1997) weighs recent exemplars more strongly and thus will be better able to capture a shift. This model is an extension of the general context model (GCM, Nosofsky, 1986), which predicts probability matching (Nosofsky, Kruschke, & McKinley, 1992). If, in a categorization task, GCM receives Category A feedback in 70 percent of the cases, it will predict Category A in 70 percent of the cases. Since EBRW includes GCM as a special case, it can also account for probability matching. However, it has a parameter that allows the theory to account for maximizing, namely the response criteria defining how much evidence the model needs before making a decision. The higher the response criterion, the more deterministic (i.e., maximizing) the choices predicted by the model will be, resulting in a greater sensitivity to differences between the two categories. Such a model would be very similar to my predictive behavior models, in which I modeled higher sensitivity with lower noise.

The results presented here could also be interpreted in the light of the RELACS model (reinforcement learning among cognitive strategies, Erev & Barron, 2005; see also

Rieskamp & Otto, 2006). Their model is able to capture a large variety of binary choice tasks. It assumes three different cognitive strategies that are involved in those tasks: Fast best reply (i.e., select the action with the highest recent payoff), case-based reasoning (i.e., choose the best action that led to the best outcome in a similar case in the past), and slow best reply (i.e., a slow learning of the strategy likely to maximize earnings). The pattern search process could be either modeled with a high proportion of case-based reasoning or with a high exploration in slow best reply. Both of these implementations would predict a deviation from maximizing, similar to increased noise in the ACT-R model I applied.

1.10.3 Conclusion

I reported counterevidence to the small sample hypothesis of correlation detection. The modeling and empirical results support the view that differences between high and low digit spans lie in differences in predictive behavior and not in differences in perception, although the whole effect only seems to exist for men. Yet, I agree with Kareev et al. (1997) that subtle benefits can follow from what are commonly seen simply as limitations (Hertwig & Todd, 2003; Schooler & Hertwig, 2005). For example, forgetting has been interpreted not simply as regrettable failures of the memory system, but as reflecting statistical patterns with which information recurs in the environment (J. R. Anderson & Schooler, 1991).

Therefore, while I am, in general, sympathetic to the idea that limitations of cognitive capacities can serve adaptive functions, I disagree with Kareev et al.'s (1997) assertion that limitations in short-term memory amplify the detection of correlation by forcing people to rely on small samples. Instead, I think that these limitations foster simpler predictive behavior, choosing the more frequent option on every trial (maximizing), due to an incapability to apply more complex predictive behavior. In this experiment, simple predictive behavior, maximizing, is more successful than any other more complex predictive behavior, such as pattern search, when the correlational structure of the task is stable. However, applying this behavior consistently seems to put people at a disadvantage if the environment changes.

To appropriately evaluate the phenomenon probability matching, I think it is necessary not only to consider how probability matching and the process I believe to underlie probability matching – the search for patterns – fare in the laboratory, but to consider how they would fare outside of the laboratory.

1.11 Probability matching reconsidered

> The remarkable thing about this is that the asymptotic behavior of the individual, even after an indefinitely large amount of learning, is not the optimal behavior… Thus, there is no indication that the individual is maximizing even in the sense of approaching a maximum. It is certainly reasonable enough from an economic point of view that he does not achieve an optimum immediately, since he does not know the situation. But it is usually assumed that after a certain amount of trial and error, the optimal behavior will in fact be found, and this reasoning is given implicitly and explicitly in most economic texts. We have here an experimental situation which is essentially of an economic nature in the sense of seeking to achieve a maximum of expected reward, and yet the individual does not in fact, at any point, even in a limit, reach the optimal behavior. (Kenneth J. Arrow, 1958, p.14).

This quote by Nobel Laureate Kenneth J. Arrow captures a rather typical, pessimistic view on the "choice anomaly" probability matching and its negative indications for human decision making.

Here, I want to step back from the specifics of the data I collected and embed the conclusions I drew from them – namely that probability matching is the result of people trying to search for patterns – into the larger picture of different approaches to probability matching. More specifically, I will argue that many of the negative conclusions about human rationality that have been drawn from probability matching stem from failing to take an ecological perspective. First, I want to briefly reiterate different approaches before I evaluate probability matching – or, more specifically, the process underlying probability matching – from an ecological perspective, stressing that a cognitive process can only be evaluated in the light of the environment in which it usually operates.

1.11.1 Different approaches to probability matching

Probability matching is commonly seen as a consistent violation of rational choice theory, because it is inconsistent with a person's goal to maximize his or her payoff. To be considered rational, proponents of rational choice theory require people to follow the principle of expected utility maximization, going back to Bernoulli and revived in the 1940s by von Neumann and Morgenstern (1947).

The cognitive limitations approach. Most people regard probability matching as one of the numerous deviations from the predictions made by expected utility theory that have been reported, all of which are usually considered as reasoning errors (Tversky & Kahneman, 1974). The explanation of those errors – including probability matching – has been echoed repeatedly: human cognitive capacities are limited (e.g., Johnson-Laird, 1983). Kahneman et al.'s (1982) heuristics-and-biases program consists of a collection of behaviors that deviate from a normative standard, which usually is either logic or probability theory. Probability matching has largely contributed to this pessimistic appraisal of human cognition. Even more so, since the conditions to behave congruently with rational choice theory are ideal – the task is simple and repeated for many trials which should enable learning.

Congruently, West and Stanovich (2003) assume that probability matching simply results from people not being smart enough to understand the task. That is, the human mind is simply not able to figure out the optimal strategy. The remedy is thus also obvious: One needs to foster the understanding of the task and to motivate people to apply maximizing to make this so-called choice anomaly disappear. Two self-evident factors in this respect are extensive learning and performance contingent payment. In general, maximizing is supported by high monetary payoffs and large numbers of trials (Vulkan, 2000). An extreme example of this is the study by Shanks, Tunney, and McCarthy (2002) in which participants were trained on up to 1800 trials and in which monetary incentives were larger than in comparable studies. Both high monetary incentives and intensive training largely boosted maximizing behavior. Note, however, that there are also studies reporting perfect probability matching despite financial incentives (Healy & Kubovy, 1981).

The repair program. Some people, in contrast to the cognitive limitations approach, have tried to keep expected utility theory not only as a prescriptive (i.e., normative) but also as a descriptive theory by making probability matching congruent with it, which could be called the repair program (cf. Gigerenzer & Selten, 2001). In order also to describe probability matching with expected utility maximization, the assumption has been introduced that different outcomes yield different utilities, even if they are equal regarding monetary payoffs. It was for example assumed that the utility for predicting the less frequent event is increased (Brackbill & Bravos, 1962). The idea behind this was that the (positive) surprise associated with correctly predicting the less frequent serves as some

kind of additional reward beyond the monetary payoff. In contrast, the utility for choosing the more frequent option over many trials is considered to be decreased due to the boredom attributed to making the same choice over and over again (Siegel & Goldstein, 1959). This is a typical example of how researchers try to resolve the discrepancy between description and prescription by amending the utility function post-hoc in an atheoretical manner without questioning the ideal of maximization or optimization (Gigerenzer & Selten, 2001). The problem with this is that by amending the utility function every behavior can be "explained" post-hoc by expected utility maximization without ever being falsifiable and without giving any insight in the processes underlying human decision making.

Searching for patterns in random sequences. The results that lower cognitive capacities actually foster maximizing instead of preventing it invited a third view: Probability matching is the result of a more complex strategy – the tendency of people to look for patterns. The results reported in this chapter (and the results I reviewed under 1.2.2) support this view. Every pattern that could possibly be correct needs to match the probabilities of the different events on the surface level.

The search for patterns indicates that people do not believe that the sequence is random. Fostering the belief in randomness increases the prevalence of maximizing. This is for example the case if the task resembled a 'gambling' task, compared to a structurally identical task that appeared to be a 'problem solving' task (Goodnow, 1955). The same occurred by giving people the opportunity to generate their own series of random events (Morse & Runquist, 1960). They were asked to repeatedly drop a rod and each time to predict whether the rod will cross a line on the floor or not when becoming stationary. It was intuitively clear to them that the sequence could not be prescheduled by the experimenter. Finally, people often have the expectation that there must be a perfect solution to the task, and this would require a pattern in the sequence. Telling them that the best possible result is to have about 75% correct answers also increased maximizing (Fantino & Esfandiari, 2002).

It is well known that people have problems to detect randomness where it exists and that they instead detect patterns even where there are none (Lopes, 1982). It is much more difficult to convince them that a sequence is random than to convince them that it is structured (Hyman & Jenkins, 1956). Is probability matching therefore just another downside of this inability to deal with randomness? On the one hand, yes.

But to conclude from this that people are irrational is premature. In the view of Herbert Simon (1990), it is important not only to look at the human mind, but also to consider the structure of the environment. This idea is captured in the concept of ecological rationality, and there have been several examples now that the structure of the environment is a sufficient explanation for many 'biases' without the necessity to assume biased cognition (Gigerenzer, 2004).

1.11.2 Probability matching reconsidered from an ecological perspective

An important cornerstone of the idea of ecological rationality is the idea that it is impossible to judge a strategy as good or bad per se. Instead, a strategy can only be good or bad given a certain structure of the environment. That is, a strategy can be good in one environment, but fail in another. The strategy that is considered to be the optimal strategy in classical probability learning studies, maximizing, is only optimal under very specific conditions. First, it is only optimal if the occurrence of an event is independent of what has occurred before (i.e., if there is conditional independence of a succession of events) and if the environment is stable. Second, it is only optimal if there are no competitors to share with. Both of these conditions are likely not to hold in many real environments.

Betting on the nonrandom structure of the environment. Outside of casinos and psychological laboratories, there are probably only few sequences of events that are indeed conditionally independent (Ayton & Fischer, 2004). If the sequence is not conditionally independent but rather systematic in some sense, it may well be worth spending some time figuring out the regularities to be able to make correct predictions on all trials. In a signal detection framework, Lopes (1982) argued that the predisposition towards patterns detection can be seen as a low criterion value to classify something as signal (i.e., generated by a nonrandom process) instead of noise (i.e., generated by a random process). Thereby, people limit the number of misses but increase the number of false alarms. Lopes further argued that often misses will be more severe than false alarms and that therefore this predisposition towards the detection of patterns is very well rational. She illustrates this argument with the fundamental attribution error, which is the predisposition of most humans to incorrectly attribute behavior of others to stable personality variables instead of attributing it to the situation. But if there is an effect that is predictable by the presence of an individual, even if only weakly so, then it is very useful to detect this.

A famous fallacy that possibly could be explained by this predisposition is the gamblers fallacy (Jarvik, 1951): people often have the inevitable impression that an event

that has already occurred repeatedly is less likely to occur again. That is, they expect negative recency. For example, in a roulette game, people think that after a series of "red" trials, "black" will be more likely on the next trial, although this is certainly false for random events. Outside of the casino, however, negative recency does exist in all cases where one samples without replacement from a finite population or from a population where replenishment is slow. Then, observing a particular outcome lowers the chances of observing that outcome again. Pinker (1997) questions the fallacious view of this phenomenon for exactly this reason: "Many events work like that. They have a characteristic life history, a changing probability of occurring over time which statisticians call a hazard function. An astute observer should commit the gambler's fallacy and try to predict the next occurrence of an event from its history so far" (p. 346). The only exceptions are devices particularly designed to fool these intuitions, such as gambling devices.

Psychological experiments are probably not deliberately designed to fool people's intuitions, but they nevertheless often do so. I think the explanation for the gambler's fallacy therefore also applies to probability matching. People try to detect patterns, which is usually useful in their natural environment. But since they are dragged out of this environment and put into an artificial situation where this pattern search fails, they appear to behave irrationally. The setting of a psychological experiment is additionally problematic in this respect, because the mere fact of participating in such an experiment is likely to raise the expectation that there will be a perfect solution to the task that needs to be figured out. Even telling people that the sequence is random does not always help because people know that deception is a common practice in psychology, which has been criticized for exactly this reason (Hertwig & Ortmann, 2001). And indeed, in many studies the sequence of events is not as random as is pretends to be, for example due to sampling without replacement or due to taking care that no event occurs more than three times in a row, both of which are structures that participants could exploit to perform better than the typical maximizing strategy (Fiorina, 1971).

Competing minds in the environment. Another important difference between the laboratory situation and the outside world is that classical probability learning experiments are individual experiments. That is, the payoff only depends on the choices of one individual and is independent of choices other people make. In the real world, however, there are competitors to share with. In such a situation, maximizing is likely to be

evolutionary unstable (Gallistel, 1990). Among a group of individuals who maximize their payoffs and choose the more frequent option on every occasion, individuals that deviate from the majority and pick the less frequent option would be naturally selected, because there are less competitors to share the outcome with. Such a countervailing selection pressure does not occur for matching. For groups of fish in a tank or ducks in a pond in both of which food was provided for instance twice as often on the one side as on the other side, it can be observed that the animals divided themselves up in the same ratio of 2:1 after only a few minutes. Thuijsman, Peleg, Amitai, and Shmida (1995) showed that simple rules could produce this ideal free distribution in an environment where there are competitors. The same rules, however, cause suboptimal matching if applied to the individual situation.

1.11.3 Conclusions

Following the tradition of Brunswik's Representative Design, the argument has been made that it is necessary not only to sample participants randomly from the population, but also to randomly sample stimuli from the environment (Dhami, Hertwig, & Hoffrage, 2004). Probability learning experiments are a typical example of unrepresentative design. People only encounter a sequence of events in which there are no sequential dependencies. Randomly sampled sequences from the environment would be likely to include a majority of nonrandom sequences, and then people would be well off searching for patterns to detect those nonrandom sequences and to detect the pattern within them. Furthermore, outside of the laboratory there are likely to be competitors, and then matching will be optimal.

This perspective on probability learning is at odds with the common view of it as irrational behavior originating in people's inability to figure out the optimal strategy. It rather demonstrates that it is likely to be an artifact of the experimental situation: People apply behavior that is usually smart to an artificial situation that is not representative of their natural environment and thereby end up behaving in way that appears to be biased. Additionally, participating in an experiment makes it even more likely that people apply behavior that does not yield the best payoff in a repeated binary choice task, because they suspect that there could be more going on than they are told, which is actually still a common practice in psychology.

Thus, probability matching can be added to the long list of choice anomalies, biases and cognitive illusions that can be explained by the structure of the environment

without blaming the cognitive limitations of the human mind. From this perspective, findings showing that maximizing is more likely for people (or animals) with reduced or lower cognitive capacities are actually no less-is-more effects. Although lower cognitive capacities are beneficial in this task, the cognitively more demanding strategy, the search for patterns, is likely to be superior outside the psychological laboratory. Not only in the laboratory, but in the real world, it often pays to explore alternatives, looking for changes and other patterns in the statistical structure of the environment. This is also supported by findings from Experiments 1 to 3 showing that people who are more likely to maximize due to a lower short-term memory capacity are put at a disadvantage as soon as the environment changes. Here I modeled what I take to be systematic exploration with random noise, but even random noise has been shown to be an effective way to escape local minima in optimization problems (Kirkpatrick, Gelatt Jr., & Vecchi, 1983).

I started by testing the small sample hypothesis, which considers less information to be helpful, and end by noting that noisy behavior, which can of course be harmful, has the potential to be beneficial. Like Kareev (2000) I emphasize that it is crucial to consider the match between a cognitive process and the environment in which it operates, because what works well in one environment may work poorly in another. No single strategy is optimal per se.

To conclude, I want to comment on an analogy that is often drawn: Many researchers studying biases and cognitive illusions claim that this will help to understand the human mind such as the study of visual illusions with artificial stimuli gives insights about the visual system. But, interestingly, no one studying visual illusions concludes from them that there is something wrong with the visual system. Taking this analogy seriously means that concluding from probability matching that people are irrational is, to borrow Pinker's (1997) words, "like calling our hands badly designed because they make it hard to get out of handcuffs" (p.346).

2 Chapter 2

Sequential Processing of Cues in Memory-Based Multi-Attribute Decisions

Although many decisions in real life depend on information we have stored in our long-term memory, relatively little experimental research has directly addressed the question how exactly attribute or cue information is integrated to form judgments and make decisions (Bröder & Schiffer, 2003b; Juslin, Olsson & Olsson, 2003) in this case. More work has investigated decision rules in environments with information supplied by the experimenter, for example on the computer screen (e.g. Betsch, Haberstroh, Glöckner, Haar, & Fiedler, 2001; Bröder, 2000; 2003; Maule, 1994; Newell & Shanks, 2003; Newell, Weston & Shanks, 2003; Payne, Bettman & Johnson, 1988; 1993). However, Gigerenzer and Todd (1999) speculated that these "inferences from givens" form an exception in everyday life which is more dependent on "inferences from memory". In addition, they argued that due to retrieval costs, inferences involving memory search would promote the use of fast and frugal heuristics.

These heuristics differ substantially from many normative theories of decision making which usually assume that more information is always better. In contrast, fast and frugal heuristics often ignore information and bet on only one good reason, which makes them psychologically plausible models of human decision making in the view of their promoters (e.g. Gigerenzer, Todd, & the ABC Research Group, 1999). Since heuristics exploit certain structures of the environment, these heuristics nevertheless do not need to be inferior to more complex decision strategies and can even outperform them under certain conditions (e.g. robustness in cross-validation, see below). The general idea behind this is that no strategy is good or bad per se, but only in relation to a certain structure of the environment (Johnson & Payne, 1985). A strategy can be called *ecologically rational* if there is a fit between the environment and the (cognitive) strategy (e.g., Gigerenzer & Todd, 1999).

An example of a fast and frugal heuristic that ignores most of the information and relies on one good reason only is Take The Best (TTB, Gigerenzer & Goldstein, 1996). TTB can be applied to compare pairs of objects on some criterion value, for example to

decide which of two German cities is larger. It is a lexicographic strategy searching information about cues sequentially, starting with the most valid cue. For example, the size of German cities can be predicted with cues such as whether a city is the capital of a state, whether it has a soccer team in the premier league or whether the city is on the Intercity train line. As soon as one cue discriminates between the objects (i.e., if it is positive for one object, but negative or unknown for the other), TTB stops searching for further information and decides based on this cue alone. That is, TTB is a one-reason decision making heuristic and is thereby noncompensatory: Less valid cues cannot change a decision based on a more valid cue, they cannot compensate a cue with a higher validity because they are not considered at all.

Despite ignoring information, TTB can outperform the compensatory multiple regression on binarized datasets when making predictions about unknown data, although multiple regression always considers all information and is a standard benchmark (Czerlinsky, Gigerenzer & Goldstein, 1999). By ignoring information, TTB is more robust than multiple regression. That is, TTB is more likely to consider only important information, which is likely to be still important in the future, while multiple regression is susceptible to random noise in the data which does not generalize – a phenomenon called overfitting.

After demonstrating that heuristics such as TTB can be successful, the question remained whether people actually use them. There is evidence that people's decisions indeed often can be described with TTB, especially if information search is constrained by high costs or time pressure (e.g. Bröder, 2000; 2003; Newell & Shanks, 2003; Newell et al., 2003; Payne, Bettman, & Johnson, 1988; 1993; Rieskamp & Hoffrage, 1999).

The assumption of sequential search of cues makes noncompensatory heuristics such as TTB different from global matching models (e.g., exemplar models), which have often been used to describe memory-based decision making (e.g., Dougherty, Gettys, & Ogden, 1999; Juslin & Persson, 2002; Mitchell & Beach, 1990). Contrary to noncompensatory heuristics, global matching models assume a simultaneous assessment of cues. While the assumption of sequential search is testable in an "inferences from givens" paradigm in which cues can be looked up on the screen (i.e., in a so-called mouselab design), this process of sequentially searching for cues is not observable in memory-based decisions.

This chapter has the goal to analyze response times as convergent evidence to substantiate that people indeed often use heuristics involving sequential search when making memory-based decisions. In this regard, I will reanalyze five experiments conducted by Bröder and Schiffer (2003b, 2006) and report one new experiment, which was necessary to disentangle two variables which were confounded in Bröder and Schiffer's experiments. Chapter 3 will then explore the complementary prescriptive question of how people could exploit features of their memory system to successfully order information, which is a crucial part of noncompensatory heuristics such as TTB.

As mentioned in the very beginning of this chapter, the application of noncompensatory heuristics was mostly studied in screen-based paradigms, although the heuristics have been proposed as memory-based heuristics. In contrast to the usual screen-based research paradigm, Bröder and Schiffer (2003b, 2006) implemented the idea of memory search in cue based decisions by introducing a cue-learning paradigm in which participants acquired knowledge about cues describing objects. The results broadly confirmed the claim that memory-based decisions are often noncompensatory and can be described by TTB for most of the participants.

However, when people use information from memory rather than from the screen, it is impossible to observe how they actually search for information. The inability to observe information search (called process tracing) poses a methodological challenge because one can only rely on outcome-based measures to decide which decision strategy someone is apparently applying. On the mere outcome level, compensatory procedures could produce decisions indistinguishable from noncompensatory strategies if the dimension weights are chosen appropriately (Martignon & Hoffrage, 2002). In addition, further evidence is needed to distinguish models assuming sequential feature processing (such as fast and frugal heuristics) from models assuming a global matching process, such as Image Theory (Mitchell & Beach, 1990), Minerva-DM (Dougherty, Gettys, & Ogden, 1999), or PROBEX (Juslin & Persson, 2002). Global matching models assume that a probe is compared to all information in memory resulting in an activation depending on the similarity of probe and stored information. Although the models differ in their details of describing the feature-based similarity match, the process appears as a simultaneous assessment rather than sequential feature comparison.

Recently, Bergert and Nosofsky (2007) analyzed response times as convergent evidence for an outcome-based strategy classification in a decision making task from

givens. I suggest that response time analyses could similarly be applied to investigate different strategies in memory based decisions. Such an analysis is also an important step, even if only a first one, to answer the call that models should ideally aim to be testable with different kinds of data (e.g., Jacobs & Grainger, 1994). I hypothesize that response times increase with the number of information pieces that have to be retrieved to make a decision, which differs for different items and/or strategies.

I will report response time analyses of 5 published experiments and one new experiment.

2.1 Reanalyzing Response Times in Bröder and Schiffer (2003b, 2006)

All experiments reported subsequently (Bröder & Schiffer, 2003b; 2006) employed a hypothetical criminal case involving 10 suspects of a murder: A famous singer was murdered near the pool, presumably by one of his former girlfriends. The participants were asked to help find the murderer. The basic idea of all the studies was to separate the acquisition of knowledge about the suspects from making decisions about them, so that knowledge had to be retrieved from memory when making decisions.

Each experiment consisted of four phases: First, in an anticipation learning paradigm, participants acquired knowledge about the individual cue patterns of 10 suspects, which differed on four cues (e.g., dog breed). Each of the cues could have three different values (e.g., Spaniel, Dalmatian, or Dachshund). A portrait and a name of a suspect appeared on the screen, and participants had to reproduce the cue values with appropriate feedback. All 10 patterns were repeated until 90% of the responses were correct, indicating a sufficiently reliable knowledge base in memory.

To prevent participants from making inferences already during learning, a cue hierarchy was established only in a second phase by informing them about the evidence (cues) witnessed at the site of crime and about its relative importance. The relative importance of the four cues (predictive cue validity) was established by telling participants how many witnesses agreed on them. For example they were told that four witnesses agreed that the suspect had a Spaniel dog, whereas only two witnesses agreed that the suspect was wearing leather trousers.

The third phase consisted of complete paired comparisons of all suspects in which participants had to decide which suspect was more likely to be the murderer. Importantly,

only the name of the suspects and their portrait were displayed. To decide between the two suspects, participants had to retrieve the cue values from memory.

After this decision phase, a final memory test assessed the stability of cue memory as a manipulation check.

The experiments differed with respect to minor procedural details (see Bröder & Schiffer, 2003b; 2006) and were in general very similar to the procedure reported in more detail for the new experiment I conducted in this respect (Experiment 6, see below).

2.1.1 Description of the strategies and response time predictions.

The strategies considered to potentially underlie the participants' decisions are TTB, Dawes's Rule (DR), Franklin's Rule (FR), and guessing. When comparing two suspects, the lexicographic TTB heuristic assumes that participants sequentially retrieve cues describing the suspects in the order of their validity. A person using TTB searches the most valid cue for both suspects first. If this cue discriminates, the person does not search further and makes a decision. Otherwise, searching for cues (in order of validity) continues until a discriminating cue is found. Therefore, the best (i.e., most valid) discriminating cue determines when TTB stops searching and decides, so that I predict a monotonic increase in response times depending on the number of cues that have to be retrieved until this best discriminating cue is found. Figure 2.1 illustrates four different item types with an increasing number of cues that have to be looked up in order of validity (according to TTB) until the best discriminating one is found.

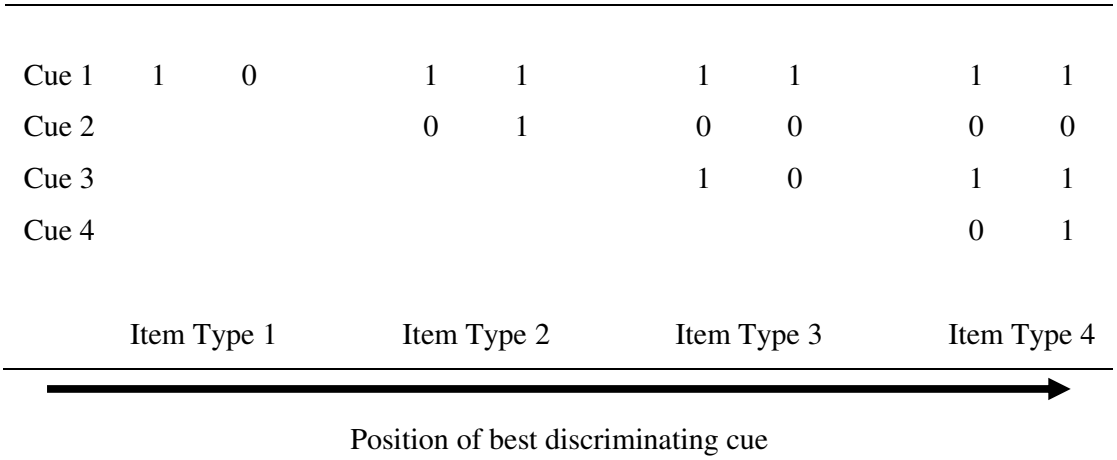

Figure 2.1. Four different item types differing with regard to the position of the best discriminating cue.

DR, going back to Robyn Dawes's (1979) work on unit-weight models, is a rule that takes all cues into account but does not consider their validity. A DR user tries to retrieve all cues and decides for the suspect that is "favored" by more cues. In case of a tie, the person has to guess. FR, similar to DR, takes all the information into account, but weighs it according to cue validity. Since less valid cues can overrule more highly valid cues, DR and FR are compensatory strategies. Both DR and FR, at least in a strict sense, require searching all cues in an unspecified order. Response times should therefore not depend on which is the best discriminating cue. Since DR and FR do not specify a search order, it is more difficult to distinguish between sequential search and global matching for users of these strategies. One prediction that, in our view, follows from the sequential search assumption but not from the global matching assumption is that people classified as using FR should be slower on average than DR users, since FR, additionally to DR, requires weighing cues according to their validity and is thus cognitively more complex.

The last strategy, guessing, consists of retrieving no cues at all and just guessing which of the suspects is more suspicious. Therefore, the response times of guessers should not vary with the position of best discriminating cue, and they should be the quickest overall.

To summarize, I expect the following qualitative pattern of response times. For participants classified as using TTB response times should increase with the number of cues TTB has to retrieve to make a decision. Such an increase should not exist for the users of other strategies. Furthermore, I expect that, on average, FR takes longer than DR, and that guessing is quicker than all other strategies.

2.1.2 Results and discussion

To decide which of the strategies someone was apparently using, the choice vector produced by each participant was classified by a maximum-likelihood procedure, details of which are provided in Bröder and Schiffer (2003a). In a nutshell, this method is designed to find the strategy that fits the data of this participant best. The method assumes a uniform response error probability and stable strategy use across trials. The likelihood of the data, given each of the models is maximized by estimating a random response error probability. The model with the highest likelihood of the data is chosen as the presumably data-generating model.

Table 2.1 contains an overview of the experiments reported previously. In sum, the results show that the need to retrieve cue information from memory induced fast and

frugal decision making, especially when cues were represented verbally and when working memory load was high.

Table 2.1: Overview of studies

Source		%TTB users x condition		N	Cue Descriptions
Bröder & Schiffer (2003b)	Exp. 1	load [a] 72.0 %	no load [a] 56.0 %	50	Blood type, cigarette brand, perfume, vehicle
	Exp. 2	memory 44.0 %	screen 20.0 %	50	Jacket, shoes, bag, vehicle
	Exp. 3	verbal 64.0 %	pictorial 64.0 %	50	Jacket, shoes, bag, vehicle
	Exp. 4	verbal 47.4 %	pictorial 26.4 %	114	dog breed, jacket, trousers, shirt color
Bröder & Schiffer (2006)	Exp. 5	verbal 69.7 %	pictorial 36.0 %	151	dog breed, jacket, trousers, shirt color
		load [a] 53.0 %	no load [a] 34.2 %		

[a] working memory load

To analyze response times, I combined participants from all 5 experiments and split them into four groups with identical strategy classifications. There were 198 TTB users, 90 FR users, 83 DR users and 44 participants who appeared to guess. Nine unclassified patterns (with identical likelihoods for more than one strategy) were excluded. For each participant, I computed the outlier-robust median response time for each of the four item types, depending on the position of the best discriminating cue. These individual response time medians were entered in the subsequent ANOVA[6].

The mean response times for each strategy group are shown in Figure 2.2. There was a main effect of the position of the best discriminating cue, Greenhouse-Geisser-corrected $F(2.53, 1041.48) = 20.63$, $p < .001$, showing increasing decision times in

[6] Note that also individual z-scores of response times and log-transformed response times were computed to control for between-subject variability and bias due to outliers. These alternative measures yielded the same patterns of significant results in all analyses.

general. There was also a main effect of strategy, $F(3, 411) = 6.92$, $p < .001$ and, much more important, a significant interaction, $F(2.53, 1041.84) = 3.41$, $p = .001$. The increase of decision times with the position of the best discriminating cue was most pronounced for TTB users. To substantiate this claim, regression slopes were computed for each individual and compared across strategy groups, showing an overall difference, $F(3,411) = 9.52$, $p < .001$. According to Scheffé posthoc tests, TTB slopes ($B = 1.31$) differed significantly from all others, whereas DR, FR, and GUESS slopes did not significantly differ from each other (B's = 0.49, 0.24, and 0.26, respectively). Neither did FR and GUESS slopes significantly differ from Zero, both $t(>42) < 1.30$, both $p > .20$, whereas TTB slopes did, $t(197) = 8.49$, $p < .001$, as did DR slopes, $t(82) = 3.27$, $p< .01$.

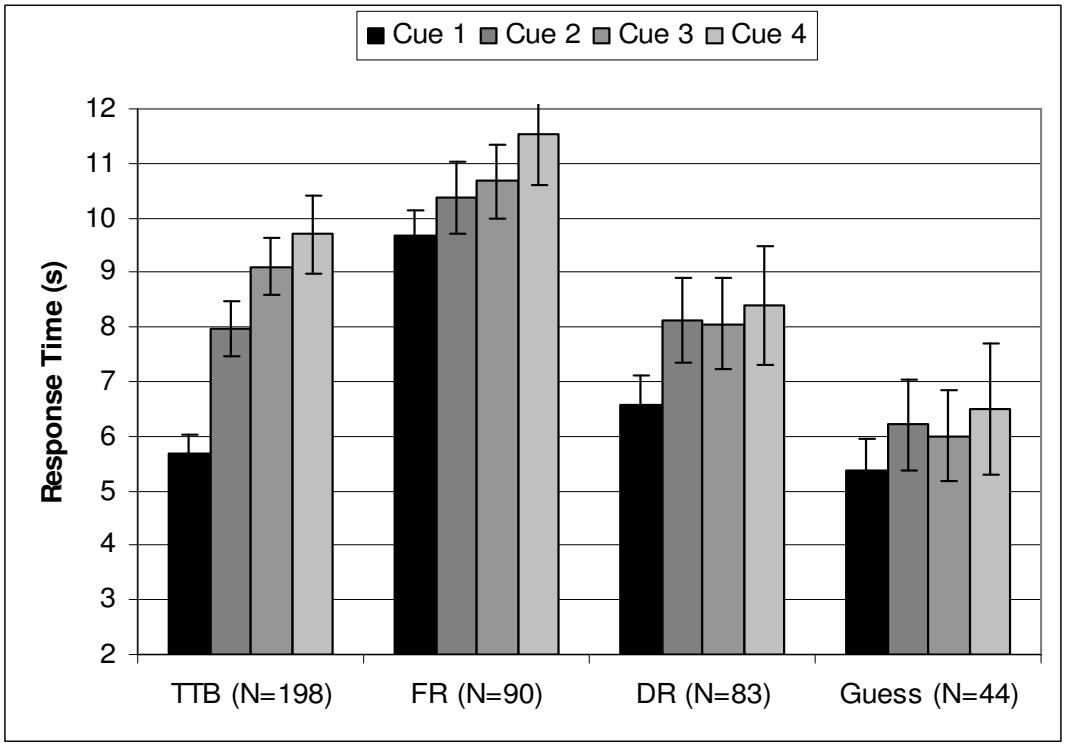

Figure 2.2. Mean decision times in seconds (and standard errors) as a function of best discriminating cue and outcome-based strategy classification on the combined data from Bröder and Schiffer (2003b, 2006).

The response times thus followed the predicted pattern and supported the assumption of sequential search: The increase was much less pronounced for FR, DR, and guessing than for TTB which would be expected if people generally search more cues than the best discriminating one, or no cues at all (when guessing). Still, there is a slight increase in response times also for FR and DR users, which I will discuss in the General

Discussion. FR users generally needed more time than DR users which was expected, given that DR users only have to count evidence, while FR users also have to weigh it. Participants with predominantly non-systematic guessing behavior generally needed less time than all others.

An alternative interpretation is that the results are not due to sequential search but to option similarity and hence item difficulty. The more nondiscriminating cues TTB has to retrieve, the more similar on average the options must be. Hence, both variables are confounded. If the sequentiality assumption is correct for TTB users, their decision times should be related more strongly to the position of the best differentiating cue than to the number of identical cue values indicating similarity or difficulty. Since position and similarity are correlated, I computed individual multiple regressions of response times on both predictors. As expected, people classified as using TTB showed a steeper increase with position (mean B = 1.08) than with similarity (B = 0.37), $t(197) = 2.76$, $p < .01$, whereas there was the opposite tendency for DR users (.14 vs. .57), $t(82) = -1.69$, $p = .09$. There was no difference in slopes for FR users (0.26 vs. -.02, $t(89) = 0.89$, $p = .37$). The same pattern of results emerges when raw rather than partial correlations are analyzed by Wilcoxon tests or t-Tests on the Fisher-Z-transformed values. Hence, there is no support for the potential alternative explanation that the TTB response time results were caused by item difficulty rather than sequential search, whereas the opposite is true for DR users.

However, a second criticism could argue that the results reported here are an artifact of a procedural peculiarity used in Experiments 1 to 5: In all cases, the cue validity hierarchy was equivalent to the order in which the cue values were learned. Hence, the seemingly sequential retrieval of cues in order of their validities might just reflect a search in order of learning. Maybe the evidence in favor of TTB and the sequential retrieval is apparent rather than real because participants only mentally repeated the learning phase and did not care about cue validities at all. To rule out this interpretation, I conducted an experiment in which I disentangled both factors.

2.2 Experiment 6

2.2.1 Methods

Design. Experiment 6 also used the paradigm of the invented murder case. A two group design was used to disentangle the two possible search orders, search by learning

and search by validity. In the *match* condition, the cue validities matched the order in which the cues were learned. In the *mismatch* condition, cue order and validity order were different. For instance, the learning order was "dog breed, jacket, trousers, shirt color", whereas the scrambled validity order of cues was "trousers, dog breed, shirt color, jacket". Hence, if the cues are numbered according to the learning sequence, the validity order in the mismatch condition was 3-1-4-2. The labels of the cues were counterbalanced by reversing validity and learning order for half of the participants (i.e., “dog” was now the most valid cue and “trousers” was the topmost cue in the learning order). Both of those cue orders were also used in two counterbalanced match conditions, in which learning and validity order coincided. Since there was no difference between the counterbalancing conditions, I merged them and subsequently only refer to cues in order of validity or in order of learning, irrespective of the actual content of the cue.

Participants. Ninety-four participants attended the study, almost 90% of them were students, about two thirds of them from the humanities. Twelve participants did not reach the learning criterion in Phase 1 within one hour (for a detailed description of the experiment, see below). Their learning phase was interrupted, but they finished the experiment like all others. Their memory performance in a final test was much worse: They only reproduced 69% of all trials correctly, while the remaining participants had a mean accuracy of 86% in this final memory task, $t(92) = 5.69$, $p < .001$. Thus, these twelve participants were excluded in all following analyses. The remaining 82 participants, 50 of them female, had a mean age of 25.6 years ($SD = 3.25$, 20 - 36). They were randomly assigned either to the match or the mismatch condition, so that there were 41 participants in each condition. Participants received 15 € for their participation with an additional chance to win 10 €, which were granted to the 5 participants with the best performance in the final memory task.

Procedure. The procedure was basically identical to Experiments 1 to 5 (Bröder & Schiffer, 2003; 2006). Still, I will describe the newly conducted experiment in detail here. First, participants were introduced to the cover story of the experiment, which was an invented criminal case: A famous singer was murdered near the pool with a tequila bottle, presumably by one of his many former girlfriends. The participants were asked to help find the murderer. In an anticipation learning paradigm, they acquired knowledge about the individual attribute patterns of 10 suspects, which differed on four cues (dog breed etc., see above). Each of the cues could have three different values (e.g., Spaniel,

Dalmatian, or Dachshund). The learning procedure was as follows: At the beginning of the trial, a color portrait of a young woman and a first name were presented. These were randomly assigned to the attribute patterns for each participant, and the 10 patterns were ordered in a different random sequence for each participant. Then participants guessed (in the first trial) or attempted to reproduce (in subsequent trials) the suspect's type of dog (or trousers, depending on condition) by choosing one of these possible cue values. Then participants received feedback: The actual type of dog appeared on the screen and, along with the participant's response, remained there for the duration of the trial. Afterwards, the suspect's type of trousers had to be guessed and so on. After reproducing the fourth cue value, the complete cue pattern of this suspect was displayed. This procedure was repeated for each of the suspects until completion of a trial on which the participant chose the correct value for all four attributes for the suspect. Hence, the whole attribute pattern of a suspect was learned before turning to the next suspect.

The procedure contained a lot of redundancy to ensure overlearning of the information. The first attribute pattern containing four cues was learned until it was reproduced without error. Then the second pattern was learned until it was also reproduced without error and so on. After the participant learned a new attribute pattern, a test followed in which all of the patterns hitherto learned were repeated. If at least 90% of the learned information was reproduced correctly in this test, a new pattern was presented or—after the 10th pattern—the learning phase was finished. If the criterion was not achieved, all patterns had to be learned again. One trial is defined as a sequence of learning all four cue values of one pattern. Even with a perfect learning strategy, participants had to complete at least 63 of these learning trials.

After this extensive learning phase, participants were told about the evidence concerning the critical cues found at the site of crime and the testimonies witnesses had given. The validity hierarchy of the four cues was established by telling participants how many witnesses agreed on them. For example they were told that four witnesses agreed that the suspect had a Spaniel dog, while there were only three witnesses agreeing that the suspect was wearing leather trousers.

In the test phase, all possible pairs of the 10 suspects (altogether 45 pairs) were presented to the participants successively in a random order. Seven items that distinguish between TTB and compensatory strategies were repeated in order to achieve a more reliable strategy assessment. Each time, they had to decide which of two presented

suspects was more likely to be the murderer. After the test phase, participants were asked again to indicate all cue values for all suspects in a final memory test. Participants needed on average 72 minutes to complete the whole experiment.

2.2.2 Results and discussion

Learning Phase. There were no differences between the match and the mismatch group with regard to the learning phase, in which I also did not expect to find a difference. Overall, people needed on average 85 trials to finish the learning phase (*SD* = 20.43; 65-142), which took them on average 36 min. The groups also did not differ in their memory performance in the final memory test. On average, people remembered 86% of cue values (89, 84, 86, 85 for cues 1, 2, 3, and 4, respectively) in this final test. That is, knowledge about the cue values belonging to different suspects was very reliable.

Strategy Classification. People were classified as users of one of the different strategies following again the outcome-based strategy classification after Bröder and Schiffer (2003a, see also section 2.1.2). In the match condition, cue validity order and learning order were identical. In the mismatch condition, however, participants may either use the validity order as a search sequence (according to TTB) or they search cues in the order of learning, which I will refer to as a "Take The First" (TTF) heuristic. Consequently, TTF was included in the set of possible strategies for the mismatch condition.

In the match condition, there were 21 TTB users, 9 FR users, 8 DR users, and 3 people who were guessing. Thus, like in the comparable Bröder & Schiffer (2003b) experiments, there was a majority of people whose decisions can best be described with TTB. In the mismatch condition, there were only 11 TTB users, 8 FR users, 12 DR users, and 5 people that appeared to guess. In addition, 5 of the 41 participants were classified as using TTF. The strategy distributions across conditions differed significantly, $\chi 2(4, N = 82) = 9.48$, $p = .05$. However, if TTF and TTB participants are combined into one class, the difference between the conditions disappears, $\chi 2(3, N = 82) = 2.04$, $p = .56$. Hence, the data support the interpretation that the proportion of TTB users found when validity and learning order are confounded may be a composite of "real" TTB users and others using TTF.

Decision times. There was no difference between the match and the mismatch condition regarding group decision times, $F(1, 80) = 0.06$, and no interaction between cue and condition, $F(2.17, 173.7) = 0.14$. Therefore, decision times for people classified as

using the same strategy were pooled across the conditions. The mean decision times for each outcome-based strategy are presented in Figure 2.3. The pattern is very similar to the one obtained in the reanalysis combining Experiments 1 to 5 (see Figure 2.2), although somewhat noisier since the sample size is only one fifth of the combined sample size of Experiments 1 to 5.

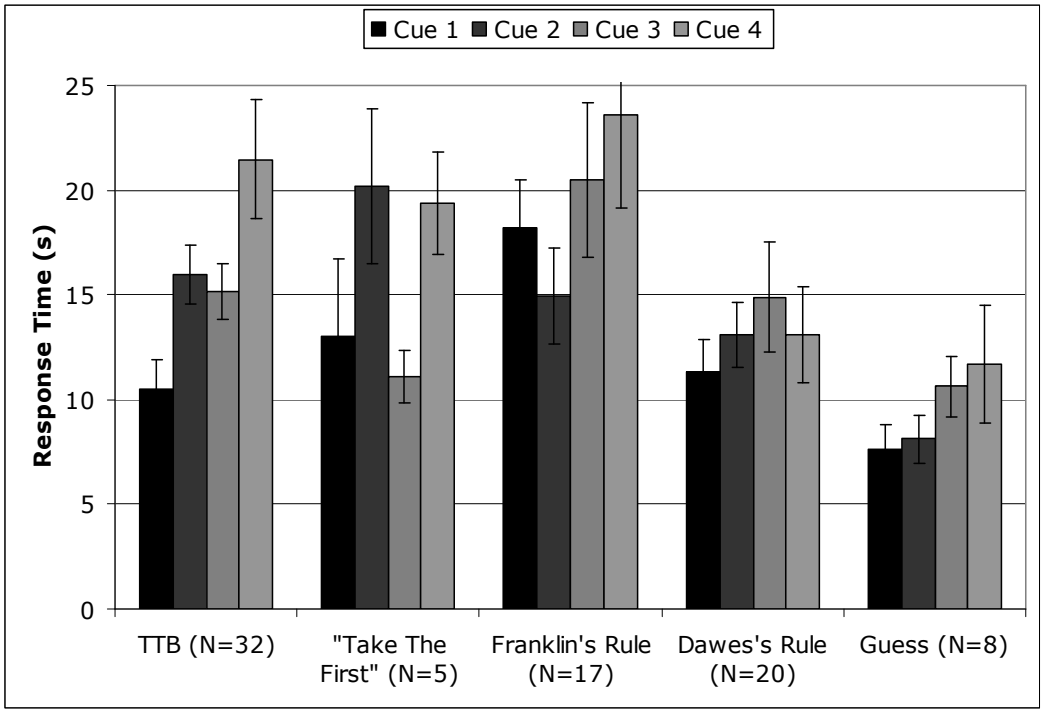

Figure 2.3. Mean decision times in seconds (and standard errors) as a function of best discriminating cue and outcome-based strategy classification in Experiment 6 including five participants who apparently searched in the order of cue retrievability ("Take The First") in the mismatch condition.

The main effect of cue is significant, $F(2.08, 160.19) = 6.55$, $p = .002$. There was also a slight effect of strategy, $F(4, 77) = 2.07$, $p = .09$, and an interaction between cue and strategy, $F(8.32, 160.19) = 2.40$, $p = .02$. Again, TTB users show the largest increase in response times, followed by FR, DR and Guessing. Response times are again generally higher for FR than for DR. A striking pattern can be seen for the five TTF users: They show the shortest response if Cue 3 discriminates, followed by Cue 1, Cue 4 and Cue 2. This pattern of response times exactly matches the learning order of the cues. Note that this result is far from trivial: The strategy classifications were exclusively based on choices. Hence, this congruence of classification and response times constitutes true converging evidence for this strategy and the notion of sequential cue search.

2.3 General Discussion

The goal of this chapter was to find converging evidence for processes assumed to underlie memory-based multi-attribute decisions. Direct process tracing is not possible for memory-based decisions. Therefore, I analyzed response times to validate the idea of sequential cue search in multi-attribute decisions from memory and as an independent source of support for the outcome-based strategy classification method.

In this manner, I have tried to answer the call that models should ideally aim to be testable with different kinds of data (e.g., Jacobs & Grainger, 1994). The current specification of the models of decision making I tested, however, only allowed for making qualitative predictions about patterns of response times, but they did not allow for making quantitative predictions or for making predictions about the distribution of response times, which is another important feature (e.g., Grainger & Jacobs, 1996; Ratcliff, 1978; Ratcliff & Smith, 2004). A promising step of specifying similar heuristic models, such as the recognition heuristic (Goldstein & Gigerenzer, 2002), has been made by Schooler and Hertwig (2005) who embedded the recognition heuristic within the ACT-R framework (e.g., J. R. Anderson et al., 2004). I believe that this would be similarly possible for the models discussed in this chapter, so that more precise predictions will be hopefully possible in the near future.

Generally, both Bröder and Schiffer's (2003b, 2006) experiments and the new Experiment 6 have shown that simple strategies are rather prevalent in memory-based decisions, especially if cognitive load is high. Both the reanalysis of the 5 published experiments and the results of the new experiment support the idea of sequential cue search. For users of TTB this support is clearest, as their response times increased with the number of cues that need to be looked up until the best discriminating cue was found. This result seems better explained by sequential search than by global matching because regression analyses revealed only a weak relation between RT and item difficulty based on feature similarity. For users of DR and FR, the increase based on the position of the best discriminating cue was much weaker and, contrary to TTB users, depended more on item difficulty based on similarity (at least for DR). In principle, this could be explained by both sequential search and global matching. However, I think that the additional finding that DR users are generally quicker than FR users is more supportive of sequential search assuming that DR users only add up information, while FR users additionally weigh it. At least, it is not clear to me how a global matching model could explain this difference.

Finally, for TTF users, response time patterns fitted the presumed search order and stopping rule, which also favors sequential search.

The strength of these sequential search effects is surprising. In none of the experiments were people instructed to decide quickly, which is usually a prerequisite to obtain interpretable response time data. Large inter- and intraindividual differences normally inflate noise and demand for large effects.

A slight increase in response times for users of compensatory strategies, as observed, is likely to occur nevertheless because it is possible that someone classified as using FR sometimes applies TTB. Moreover, retrieving information in order of validity also makes sense for compensatory strategies. For example, a FR user who knows that the two most valid cues point towards one suspect does not need to look up further information because the less valid cues could never overrule this judgment.

Experiment 6 was necessary because the previous experiments confounded learning order and validity order. Therefore, it is possible that participants in the test phase did not try to retrieve cues in order of validity, but simply in the order in which they learned them. Thus, the response time pattern I assumed to support the idea of sequential cue retrieval in order of validity could have been an artifact of the experimental design. However, even in the mismatch condition of the new experiment the data support sequential cue retrieval in order of validity for many participants. That is, although it is potentially more difficult to retrieve the cues in order of validity, response times still increase with the number of the first discriminating cue, starting with the most valid one. However, the classification data reveal that a proportion of apparent TTB users actually switch to another retrieval order if this is easier to accomplish. Hence, even with "one reason decision making", people may actually consider a variety of search orders that go beyond the criteria mentioned for example by Martignon and Hoffrage (2002) and Newell, Rakow, Weston, and Shanks (2004).

This result adds a new facet to the debate about how the structure of the environment affects human decision making. Usually, the structure of the environment is discussed in relation to the performance of different decision strategies. For example, it has been shown that people are able to adapt their strategies to the feedback structure of the environment (Bröder, 2000; 2003; Payne et al., 1993; Rieskamp & Otto, 2006). That is, they learn to apply different strategies depending on which strategy is more accurate. In the current experiment there was no accuracy criterion, but strategy use tended to vary

with the difficulty of applying a strategy given a certain order of cues in the learning phase. The fit between the (learning) environment and the "applicability" of a strategy apparently matters. Newell et al. (2004) also provided evidence that people do not solely search cues in order of validity, but that instead their search order is, roughly speaking, influenced as well by validity as by how often a cue is applicable. In the study reported here, the ease of application depended on the learning order of cues and thereby on the difficulty to retrieve cues in a certain order. Outside of the laboratory, it will probably more strongly depend on the frequency of encountering cues. Our memory system is organized in a way which facilitates the retrieval of information which is recent or frequent (J. R. Anderson & Schooler, 1991; Deese, 1960; Geoff et al., 2003).

In Chapter 3, I investigate whether people can have both easily applicable and successful strategies outside of the laboratory because they can exploit their memory system which mirrors the nonrandom distribution of information.

3 Chapter 3

An Ecological Approach to Memory-Based Heuristics

The question of how people actually search for information in the real world has raised quite some interest. Juslin and Persson (2002) argued that TTB is not so simple, since all the difficult computations go into ordering cues by validity. Only given this validity order, however, can TTB be successful and simple at the same time. Dieckmann and Todd (2004) have shown that people do not seem to order cues by validities they could learn on a trial-by-trial basis. Instead, people often seem to use much simpler rules to order cues, which results in strategies that are more frugal than TTB (i.e., need fewer pieces of information) and are better than strategies using a random cue order. However, they fall behind the performance of TTB.

Chapter 2 has demonstrated that people often rely on simple cue-based strategies that process information sequentially in memory-based decisions. Building on Dieckmann and Todd's (2004) results that people seem to use simple strategies to order cues, I want to explore the possibility that people could exploit their memory system in a way that facilitates ordering cues and at the same time is successful. In particular, I want to address the question of whether the speed of retrieving information could be used successfully to order cues.

The speed of retrieving information is mostly a function of frequency and recency of encountering this information (J. R. Anderson & Schooler, 1991), and also the context of the information is important (Schooler & Anderson, 1997). That is, people are most likely to come up with cues quickly that they have encountered frequently and recently. Betting on the speed of retrieval could exploit some peculiarities of the environment. Goldstein and Gigerenzer (2002) reported that it is more likely to encounter information about objects with large criterion values: A newspaper analysis revealed that large objects (here: German cities) were mentioned more often than small objects. Moreover, the more often the name of a city appeared in the newspapers, the more likely people were to recognize its name. Building on these two correlations – the correlation between the size of a city and how often it is mentioned in the media, and the correlation between how often it is mentioned in the media and how often it will be recognized – Goldstein and

Gigerenzer demonstrated how people could successfully rely on their recognition if they wanted to predict which of two cities has more inhabitants. They labeled using recognition to infer values on a certain criterion (like city size) the "recognition heuristic".

Recognition heuristic: 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 (Goldstein & Gigerenzer, 2002, p. 76).

Schooler and Hertwig (2005), building on the recognition heuristic, have used the ACT-R framework (e.g., J. R. Anderson et al., 2004; see also Chapter 1) to specify the *fluency heuristic* (e.g., Jacoby & Dallas, 1981), which stems from the idea that fluency of reprocessing can be used as a cue in inferential judgment. In that sense, the fluency heuristic is similar to the *availability heuristic* (Tversky & Kahneman, 1973), which assumes that people use the ease of retrieving instances and the frequency of instances they retrieve to assess the probability and the frequency of events. In the view of Tversky and Kahneman, the availability heuristic can explain all kinds of biases and cognitive illusions. Others (e.g., Gigerenzer, 1996), however, have criticized the availability heuristic because its underlying processes are still not precisely specified, even more than 25 years after it has been introduced. The fluency heuristic as defined by Schooler and Hertwig transcends this criticism as it is precisely defined in the ACT-R framework and thus makes testable predictions where it will succeed and where it will fail. Contrary to Tversky and Kahneman, Schooler and Hertwig do not focus on biases, but believe that their "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" (p. 626).

Schooler and Hertwig applied the notion of fluency to the city paired comparison task, as described above. Their specification of the fluency heuristic captures both the binary connotation of recognition and the conceptualization of recognition as something that gradually increases with repeated exposure. That is, it is applicable in situations in which one object is recognized while the other is not as well as in situations in which both objects are recognized, but one of them is recognized faster because it has been encountered more often.

According to Schooler and Hertwig's (2005) model of the fluency heuristic for paired comparisons, an attempt is made to retrieve both representations of objects from memory. If an object's representation can be successfully retrieved, then it is recognized. But even if both objects' representations can be retrieved, it is still likely that one will be retrieved faster than the other, which is the notion of fluency. Fluency is related to the number of exposures and is therefore informative. For example, it is likely that one has heard more often of "Berlin" than of "Greifswald". Even if someone recognizes both of them, "Berlin" is likely to be recognized faster, which can be used to infer that it is the larger city.

Here, I am interested in extending the idea of fluency to the level of cues. More specifically, I believe that not only the fluency of recognizing an object is informative, but that the order with which cues about the object come to mind (the more fluent, the quicker) could be useful. If, in the German cities task, someone recognized both objects, and wanted to rely on further cues to predict which of the two has more inhabitants (which I will simply refer to as "city size" subsequently, even when referring to the number of inhabitants), the memory system could help people in the following ways.

People may not only encounter the names of these cities more frequently (which was the level of analysis in Goldstein and Gigerenzer, 2002), but are also likely to encounter more information about these cities. Encountering more information means two things in this context. First, people may encounter more different pieces of information about larger objects and, second, they may encounter each piece of information more frequently. Note that this will hold especially for positive cue values, that is, for the presence of cues (e.g., that a city has an airport, a train station, etc.). Therefore, the fluency of retrieving positive cue values could be positively related to the size of the object (i.e., city size), and subsequently, the order of retrieving positive cue values will be informative because more positive cue values will be available more quickly about the larger object.

For negative cue values, this relation might be the exact opposite. For example, one might know fairly quickly that the rather small city Moers does not have a soccer team. However, to know that the 7th largest German city, Essen, also does not have a soccer team, might take longer. The reasoning behind this is that people will know more positive cue values in general about the larger city Essen, and these other positive cues could make them believe that Essen is more likely also to have a soccer team. Although this is not the case, people might still think more about it. For Moers, in contrast, many

people probably do not even have a single positive cue value in mind and thus can also answer quickly that it does not have a soccer team. That is, the fluency of negative cue values could be negatively related to city size (i.e., be retrieved more quickly for smaller cities), which again is informative.

A retrieval-based strategy that attempts to predict which of two objects is larger could exploit the positive relation between city size and the fluency of positive cue values by betting on the object about which more positive cue values come to mind more quickly. Such a strategy could also exploit the negative relation between city size and the fluency of negative cue values by betting against the object about which more negative cue values come to mind more quickly. Both of these features could fuel the success of retrieval-based strategies.

A further feature of the memory system that a retrieval-based strategy could exploit is that correct cue knowledge should come to mind more quickly than incorrect cue knowledge, which is a common finding in the memory literature (e.g., Ratcliff & Smith, 2004). This is plausible since people will almost exclusively encounter correct instantiations of object-cue relations in the environment. Therefore, a person will be more likely to end up with a wrong cue value if he or she makes an effort to think about a particular cue value that does not come to mind quickly, as is the case with strategies that enforce a particular order on the individual (such as TTB). In contrast, retrieval-based strategies should be fairly robust against incorrect cue knowledge, because incorrect cue values will only be retrieved very slowly and thereby often ignored.

An experiment was designed to study whether these features exist in a real world environment. These data will then be used to simulate the accuracy of retrieval-based strategies building on these features in comparison to other strategies.

3.1 Experiment

The approach I will take to explore whether ordering cues by fluency is ecologically rational relies on the assumption that people's memory in some sense mirrors the environment, as was shown by Goldstein and Gigerenzer (2002), where recognition rates mirrored environmental frequencies. Therefore, I believe it to be appropriate to investigate the knowledge people have stored in memory as a reflection of the environment instead of analyzing the environment directly. More specifically, instead of analyzing how often people would actually encounter certain object-cue combinations

(such as Hamburg – airport) in the environment, I want to study directly how quickly people can retrieve such object-cue combinations. The basic idea of how this can be done is to ask people questions of the type "Does object X have cue Y?" The assumption here is that the faster people will be able to answer this question, the more fluently available this object-cue knowledge is to them. Therefore, an experiment was designed to assess the fluency of cues for different objects. These fluencies provided by the data will then be used to order cues and to simulate the performances of different cue-based strategies that apply this order. These strategies will be called retrieval-based strategies and will be compared to other models.

The experiment needs to assess people's knowledge about real cues and real objects because the ecological argument I am making relies on a nonrandom distribution of cues and objects in the real world, which, in turn, are reflected in different fluencies. The selection of the real world environment to be studied is thus crucial. First, it is important to select an environment about which people have some kind of knowledge. Second, it is important that the cues that describe the objects within the environment are cues that people would probably use were they to make an inference about the objects. And finally, the cues need to be defined and assessed precisely so that their relation to the criterion can be specified. In the next section, I will describe the selection of the environment.

3.1.1 The environment

The environment in which I studied how the fluency of cues could be implemented successfully in cue-based strategies was German cities. German cities were already studied to assess the performance of different heuristics in comparison to other models by Gigerenzer and Goldstein (1996). The German city environment I studied was an updated and altered version of Gigerenzer and Goldstein's environment. Most importantly, it was smaller, so that it was possible to ask participants exhaustively about all city-cue combinations. Furthermore, the selection of the cues was more strongly based on what people believe to be important cues, resulting in a somewhat different, but still overlapping selection of cues. In the following, I want to describe how the cues and the cities were selected.

Selection of the cities. The cities included in the study I will describe below were selected based on the publicly accessible Urban Audit database on the internet (Urban Audit, 2006a), a large database of European cities which provides much information about

each city. The most recent data accessible in this database is from 2001. Urban Audit aims at a balanced and representative sample of cities in Europe. To obtain such a selection, a few simple rules were followed (quoted from Urban Audit, 2006b):

- Approximately 20% of the national population should be covered by the Urban Audit.
- All capital cities were included.
- Where possible, regional capitals were included.
- Both large (more than 250.000 inhabitants) and medium-sized cities (minimum 50.000 and maximum 250.000 inhabitants) were included.
- The selected cities should be geographically dispersed within each Member State.

Among data for 258 European cities, the database provides information about 35 German cities, ranging from size, demography, and social aspects to economic indicators. Out of these 35 cities, I selected 20 cities at random while at the same time ensuring that the relation between cues and city size, described by cue validity (see below), did not differ between the set of 35 cities and the set of 20 cities. This procedure guaranteed that the subset of 20 cities reflects the Urban Audit set of 35 German cities well. Furthermore, I ensured that the set of 20 cities did not contain both the cities Frankfurt (Main) and Frankfurt (Oder) because this could be confusing for participants. The cities in the final set were: Berlin, Munich, Cologne, Frankfurt (Main), Essen, Bremen, Leipzig, Nuremberg, Dresden, Wuppertal, Bielefeld, Wiesbaden, Augsburg, Freiburg, Erfurt, Mainz, Göttingen, Moers, Trier, and Weimar. The most recent information on city sizes was dated to December 31st, 2003 (City Population, 2006).

Selection of the cues. For the selection of the cues, two criteria were important. First, the cues needed to reflect information that people actually consider to be important to predict city size. Second, the cues needed to be precisely definable so that it was possible to assess the cue values for different cities. Pachur, Bröder, and Marewski (2006) kindly provided data on which cues people actually believed to be important. They asked 100 participants in an open format which information they believed to be predictive of city size and categorized those answers. Among the most frequently mentioned categories, I selected those cues I believed to be precisely definable.

Note that all cues are binary, most of them naturally so, and the others were dichotomized at the median. A value of one points in the direction of the city being large, while the value of zero points in the direction of the city being small. Based on these binary cue values, the relation between the cues and the criterion (here: city size) can be described by the cue validity. The validity of a cue measures how often one cue alone would successfully predict which of two cities is the larger one in all possible pairs of cities, given that the cue discriminates. That is, to compute the cue validity, one first needs to generate all 190 paired comparisons from the 20 cities. Then, on all the paired comparisons on which the cue discriminates (i.e., is 1 for one city and 0 for the other), one counts how often the larger city actually has the larger cue value. More specifically, the cue validity is defined as the number of correct predictions divided by the number of all predictions where the cue discriminates between the objects.

The discrimination rate of a cue is the relative frequency with which a cue discriminates between pairs of objects. The more often a cue discriminates between pairs of objects, the more useful it is. On the other hand, a cue with a high ecological validity and a very low discrimination rate might not be all too useful. For example, the National Capital cue in the German city set has a validity of 1, since Berlin is the capital and the largest city within Germany. However, the cue only applies to comparisons between Berlin and other cities and is useless in all other comparisons between these other cities.

The cues are described in table 3.1. Appendix A lists a more detailed description of where the data was taken from and how the cue values were exactly determined for each of the cities (Appendix A, Table A). It also provides the complete German city environment as used for the study (Appendix A, Table B).

Table 3.1. Definitions of the cues

Cue	Validity	Discr. Rate	Description
National Capital	1.00	0.10	Is the city the national capital?
Expositions	0.91	0.52	Are there international expositions in the city?
Airport	0.90	0.52	Does the city have an airport?
Train Station	0.89	0.10	Does the city have a train station where long-distance trains stop?
Soccer Team	0.84	0.51	Does the city have a soccer team in the premiere league?
Industry	0.84	0.51	Is there important industry in the city?
University	0.75	0.19	Does the city have a university?
Harbor	0.71	0.52	Does the city have an important harbor?
Infrastructure	0.66	0.51	Does the city have an above-average infrastructure?
State Capital	0.64	0.48	Is the city the capital of one of the 16 federal states?
Tourism	0.61	0.51	Is the city visited by an above-average number of tourists per inhabitant?

3.1.2 Methods

Design and Procedure. As mentioned before, the main purpose of the study was to assess the accuracy of people's cue knowledge and to assess how fluently they retrieve the cues. The cue values people believe the cities to have and their fluency will then be used to simulate how good sequential decision strategies can be by ordering cues according to their fluency. Although therefore the cue knowledge was the main interest of the study, I additionally assessed other aspects of people's knowledge about the cities.

The study consisted of four parts: a recognition test, a cue knowledge test, an inference test, and a questionnaire about criterion knowledge, which were presented in this order to all of the participants. In the introduction, participants were informed that the purpose of the study is to find out what people know about German cities. In the recognition test that followed, participants were asked for each of the 20 cities whether they had seen or heard the name of the city before. The cities were presented in random order. Note that the purpose of the recognition test was mainly to control whether some knowledge about the cities could be expected, which would not be the case if many people did not even recognize the city names. My expectation was that a majority of people would recognize all of the city names.

In the cue knowledge task, participants first learned about the eleven cues described above. They were given precise definitions of what the cues mean. If the cue was dichotomized at the median, they were told that having a positive cue value (e.g., having a good infrastructure) here means to have a higher value than the average German city. Furthermore, they were again given an overview of all the 20 cities to make the reference class clear to them. Then, participants were asked about each of the 220 city-cue combinations in the format "Does city X have cue Y?" The order of these city-cue combinations were randomized, so that participants could neither anticipate on which city nor on which cue they would be probed on the next trial. Furthermore, participants were instructed to answer as quickly as possible, but without making avoidable mistakes. Note that they could only answer yes or no. That is, if they did not know the cue value, they were forced to guess. The reason for that was that response times are likely to be cleaner if there are only two options, one for a finger of each hand. Including a third option ("do not know") could have increased the noise on the response time data.

The inference task consisted of complete paired comparisons between the 20 cities. That is, there was a total of 190 paired comparisons, in random order. On each

comparison, participants had to indicate which of the two cities was the larger one. This task was included to have people's accuracy on that task as a benchmark for the simulations which will be described later.

Finally, a questionnaire was given to participants to assess direct criterion knowledge. That is, participants were asked to indicate the number of inhabitants for those cities where they approximately knew it. Assessing people's direct knowledge of the criterion (the size of the cities) was important because it was possible in principle that their criterion knowledge was better than their cue knowledge. If this were the case, they could use their knowledge about city size in order to estimate the cue values, which could pose a problem since I want to model the exact opposite (i.e., I want to use their cue knowledge to build models inferring city size).

Participants. Forty-two participants (29 female) participated in the study. Their mean age was 25 years. The experiment lasted on average about one hour plus the time people spent on the questionnaire. Participants were paid 0.05 € for each correct answer in the cue knowledge and the inference task, and they earned on average about 16 €.

3.1.3 Results and discussion

Here I want to report global results and analyze which features of the cue knowledge retrieval-based strategies could potentially exploit to make good predictions. As expected, it was the case that people's recognition of the cities was very high. The median recognition rate was 95% (i.e., 19 out of 20 cities). Only seven participants failed to recognize more than one city. Furthermore, their knowledge of cue values was quite good. On average, their cue knowledge was correct on 74.4% of the city-cue combination questions (SD = 4.1%). In the inference task, they correctly predicted which of two cities was larger in 79.9% of the paired comparisons (SD = 5.5%). Cue knowledge was positively correlated with the accuracy on the inference task, $r(40) = .41$, $p = .008$.

The criterion knowledge as assessed on the questionnaire was very low. Out of the 20 cities, the median number of cities participants specified some correct criterion knowledge about was only one, and that was Berlin in almost all cases. This suggests that criterion knowledge did not play an important role overall.

In the introduction of this chapter I speculated about different ways a retrieval-based strategy could successfully exploit the fluency of cues, which will be explored with global response time analyses in the following. Note that all subsequent analyses deal with

the common (i.e., base 10) logarithm of the city sizes, since the distribution of city sizes is highly skewed.

I have speculated that the fluency of cue values could be a function of the size of the city because one is more likely to encounter more information about larger cities. For each city, the median response time is computed on all participants' response times generated on answers to questions about city-cue combinations of this city, separately for all answers, positive answers (i.e., indicating the belief of the presence of a certain cue), and negative answers (i.e., indicating the belief of the absence of a certain cue). As hypothesized, there was a correlation between city size and the mean logarithmized response time of all answers on questions about cue values concerning this city, $r(18) = -.446$, $p = .049$. That is, the larger the city, the more fluent people's cue knowledge about the city. I have speculated additionally that this correlation should exist mainly for positive answers, while it should be reversed for negative cue values.

As hypothesized, there was an even stronger correlation between city size and the mean logarithmized response time of positive answers, $r(18) = -.673$, $p = .001$. The opposite is true for negative answers: the correlation between city size and the mean logarithmized response time of negative answers indicates slower answers for larger cities, $r(18) = .615$, $p = .004$. Figure 3.1 depicts the scatter plots of the relation between city size and the median response times of A) positive and B) negative answers.

Another potential feature to be exploited by retrieval-based strategies is that the retrieval of correct cue values (i.e., correct answers to questions about city-cue combinations) might be faster than the retrieval of incorrect cue values (i.e., incorrect answers to questions about city-cue combinations). By relying on retrieval speed to order cues, retrieval-based strategies could be prevented from considering incorrect cue values. To test whether correct answers were indeed faster than incorrect ones, I computed the median response time on correct and incorrect answers for each participant. Those response times were then analyzed with a within-subjects t-test across all participants. As hypothesized, the median response times on incorrect answers were slower on average ($M = 2173$ ms; $SD = 593$ ms) than on correct answers ($M = 1758$ ms; $SD = 347$ ms), $t(41) = 8.48$, $p < .001$.

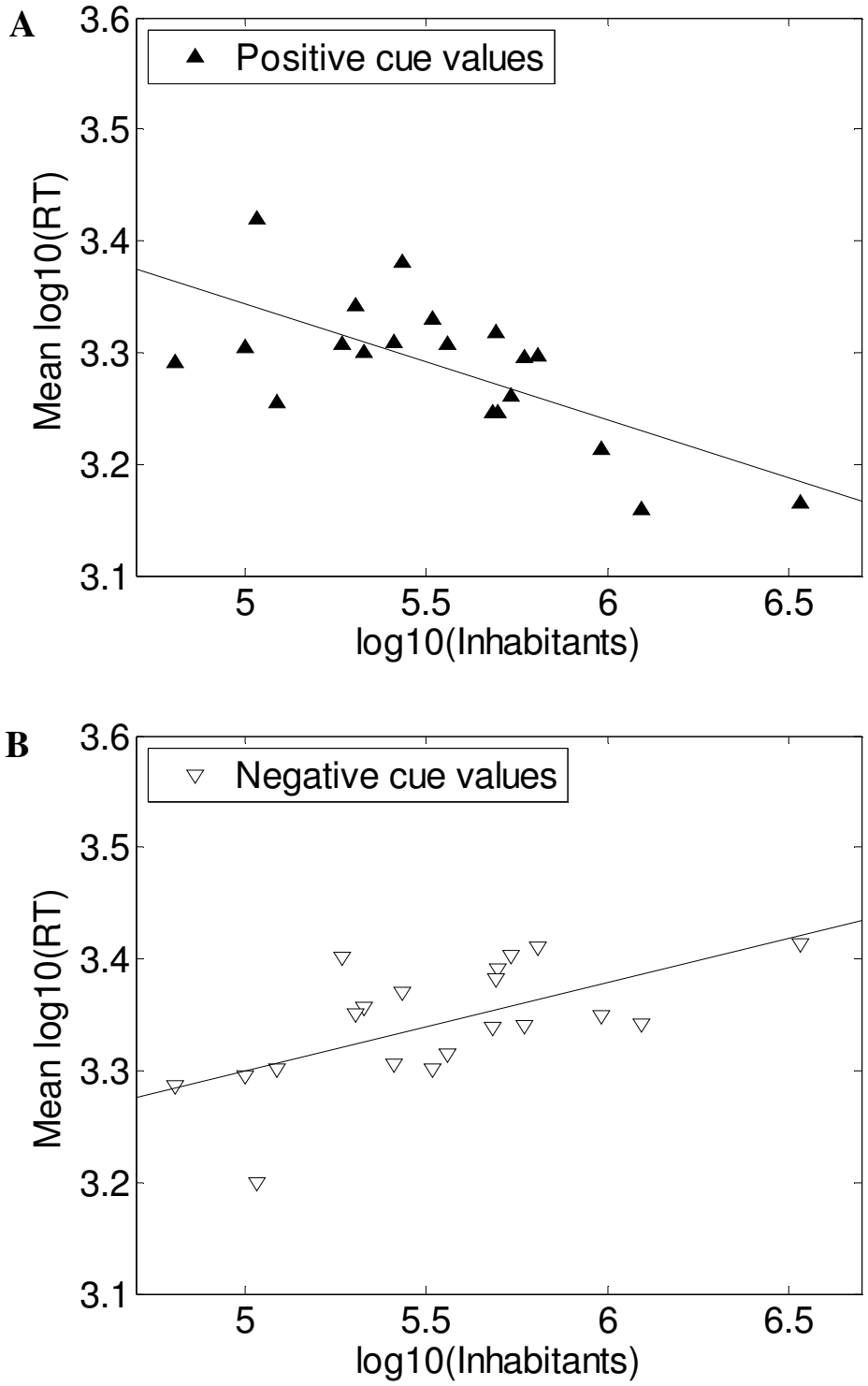

Figure 3.1. The relation between the common (i.e., base 10) logarithm of the cities' inhabitants and the mean logarithmized response times (across all participants) of A) positive answers, and B) negative answers. The line is the least squares line.

3.2 Simulating the Performances of Retrieval-based Strategies

In this section, I want to use the cue fluencies as assessed in the Experiment to simulate the performance of retrieval-based strategies which use the cue fluencies to order cues. That is, the nature of this section is rather a prescriptive than a descriptive one. It attempts to answer the question of how successful people would be when relying on cue fluency to order cues, not the question of whether people actually do so.

In addition to the participants' performance on the inference task, two other types of benchmarks will be considered against which to test the performance of retrieval-based strategies. The first one is the performance of other strategies in the very same environment. The strategies I consider here are Minimalist, TTB, and multiple regression, which will be described below. The second benchmark will consist of comparing strategies relying on retrieval order to strategies that are identical except that they ignore the retrieval order and search for information in random order. This benchmark allows for the assessment of how much a strategy gains with regard to performance by assuming a search in order of retrieval. In the following section, the competing strategies will be described before turning to a description of retrieval-based strategies.

3.2.1 The competitors

Minimalist. On each paired comparison, this strategy searches cue values on one cue for both objects in random order. That is, it assumes a cue-wise comparison of objects by always comparing objects on the same cue. As soon as a cue is found that discriminates between the two objects, Minimalist stops searching and settles for the object favored by this cue. That is, Minimalist is similar to TTB in that it is noncompensatory, but it differs with regard to the search order.

TTB. This strategy searches cue values on one cue for both objects in order of validity, starting with the most valid cue. Thus, it also assumes that objects are compared cue-wise. As soon as a cue is found that discriminates between the two objects, TTB stops searching and predicts that the object favored by this cue has the higher value on the criterion.

Multiple regression. As the most demanding strategy in the set, multiple regression always takes all cues into account and linearly combines them to predict the criterion. It assigns weights to the different cues, which also take dependencies between cues into account, to minimize the least squares difference between a predicted and the true criterion

value. That is, it makes continuous predictions. The performance on the paired comparison task was assessed by counting on what proportion of the pairs the city which is predicted to be larger is also actually the larger one.

After having described the competitor models, I now want to describe one class of retrieval-based strategies as an example,before then comparing the performances of these different strategies.

3.2.2 DifferX – A class of retrieval-based strategies

In total, I have studied three different classes of retrieval-based strategies. Since all of them are rather similar and also yield similar results, I will focus only on the most successful one here, for the sake of simplicity. The other two strategy classes and their results are reported in Appendix B.

The common idea behind all of these strategies is that information is looked up in the order in which it is retrieved. That is, for each paired comparison, there is one vector generated consisting of 22 pieces of information (2 cities x 11 cues). These 22 cue values are then ordered by the response time as assessed from each participant on the cue knowledge test, starting with the shortest response time. Note that this results in an important difference to cue-based strategies such as TTB: While TTB assumes that objects are compared on each cue in a lexicographic order, the retrieval strategies dispense with a cue-wise search but instead gather evidence for and against each object until a threshold is reached. That is, the retrieval-based strategies assume that single pieces of information (i.e., cue values for one object) are retrieved sequentially so that objects are not necessarily compared on the same cues. For example, someone could quickly retrieve the fact that Hamburg has an airport and that Heidelberg does not have a soccer team in the premier league, and use this information to infer that Hamburg is the larger city. Furthermore, it is entirely possible that someone would retrieve plenty of cues for Hamburg, while at the same time nothing comes to his or her mind about Heidelberg, allowing for the same inference.

This also has consequences for measures of frugality describing how many cues need to be looked up according to a strategy in order to make a decision. In Gigerenzer et al. (1999), looking up one cue for both objects was counted as one cue that was looked up. In the following, this is counted as two pieces of information because one cue value is looked up for each of the two objects. This is important to keep in mind when comparing

the frugality of the retrieval-based strategies to the frugality of the heuristics described in Gigerenzer et al. and related papers.

Note that in the class of strategies described here (and also in the other strategies described in Appendix B), negative information about one object counts as information favoring the other object. That is, negative information about object B favors object A identically, as does positive information about object A.

The class of strategies I will describe in more detail is labeled *DifferX*. These strategies require a fixed difference in pieces of favoring information between the two objects before they stop searching and decide. That is, this strategy continues to look up information until this information difference is reached. The X in the label DifferX specifies the decision threshold (here: the difference required to stop search). A Differ2 strategy, for example, requires that the difference between the objects A and B with regard to the number of pieces of favoring information is two. In this example, the difference would be two given one positive piece of information about object A and one negative piece of information about object B. The difference would also be two with two positive pieces of information for object A, or with two negative pieces of information for object B. Should all pieces of information be looked up before the decision threshold is reached, the strategy settles for the object that is favored by more pieces of information. Should there be a tie, the strategy guesses.

3.2.3 Methods

All of the strategies were used to predict which of two cities is larger on the 190 paired comparisons which were also presented to participants in the Experiment. The models could rely on up to all of the 11 cues (see table 3.1) per city, depending on their stopping rule.

An important distinction needs to be made here between ecological cue values and individual cue values. The ecological cue values refer to the true cue values in the environment (i.e., those reported in Appendix A, Table B). The individual cue values refer to the cue values as indicated by the 42 participants in the Experiment. Therefore, models using individual cue values were simulated separately for the 42 participants and then averaged across them.

Models were tested both on the ecological cue values as an upper benchmark and on the individual cue values for each participant separately. When relying on the individual cue values, for TTB and multiple regression it is additionally important to

differentiate how they rely on them (i.e., how they order or weight them). The order of cues for TTB and the cue weights for multiple regression were first determined on the ecological cue values, and then the strategies could apply the individual cue values using the ecological cue order or the ecological cue weights to make predictions (which would be a type of crossvalidation). Alternatively, the order of cues for TTB and the cue weights for multiple regression were determined on each participant's individual cue values separately, which is a more extensive fitting procedure. Then, the strategies could apply the individual cue values using the individual cue order or the individual cue weights to make predictions.

For Minimalist, there were 100 different cue orders randomly determined, separately for each of the 42 participants. This was done similarly for the random cue order strategies that are counterparts of the retrieval-based strategies (i.e., that are identical to the retrieval-based strategies except that they ignore the retrieval order and use a random order instead): There were 100 different cue orders simulated randomly, separately for each of the decision thresholds, for each of the pairs and for each of the 42 participants.

3.2.4 Results

The competitors' performances. Performance of the competitor strategies on the city environment is depicted in Figure 3.2. Note that performance using ecological cue values and performance using individual cue values are also included. Furthermore, for TTB and multiple regression, if individual cue values are used, there is a further differentiation depending on whether they are ordered or weighted based on ecological cue values or based on each participant's individual cue values. The participants' average performance on the inference task is depicted as well.

The frugality of the different strategies is also included in the figure. Note that the frugality of a strategy did not depend on whether the strategy relied on ecological or on individual cue values, except for very minor differences. Hence, only one frugality measure per strategy is reported. Further note that this measure of frugality is slightly different from the frugality measures reported in Gigerenzer et al. (1999). While Gigerenzer et al. counted it as one if one cue was looked up for both objects, I count this here as two, since two different pieces of information have been looked up – one cue value for object A and one cue value for object B. This is important in comparing the

competitors' frugalities with the retrieval-based strategies, because those dispense with the idea of always comparing objects on the same cue (see section 3.2.2).

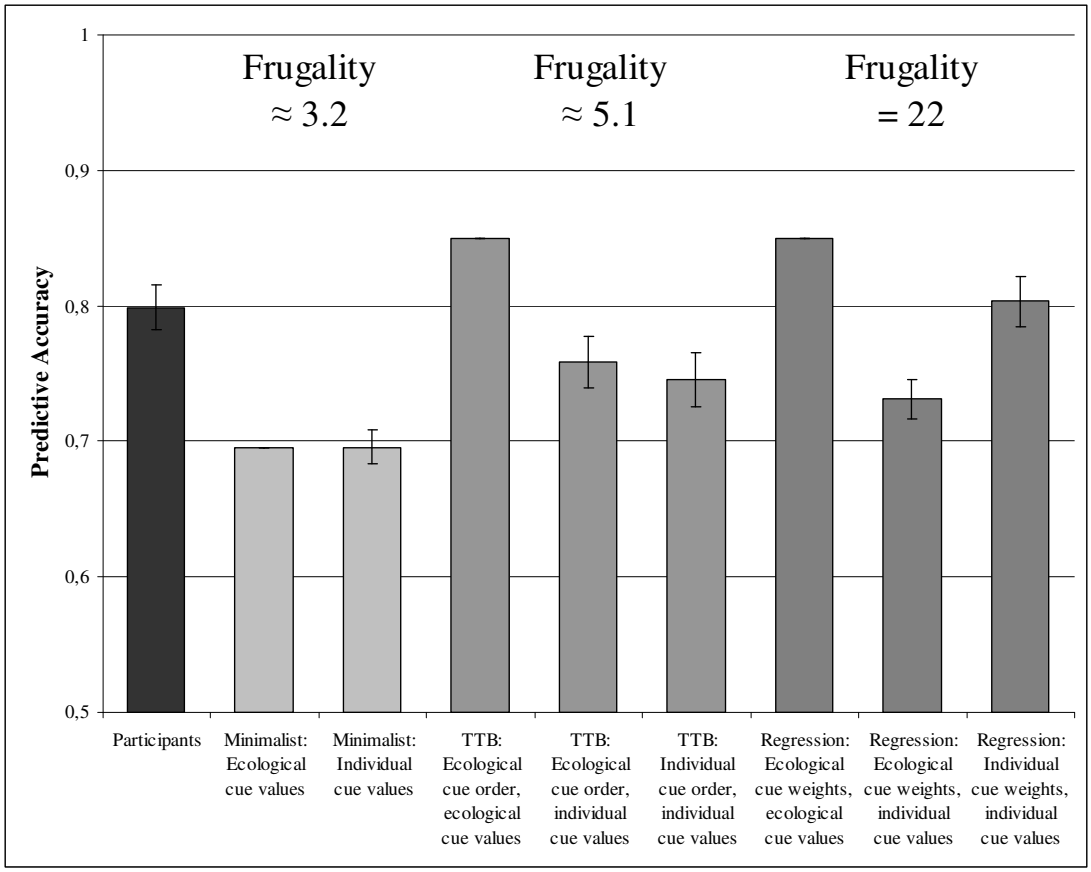

Figure 3.2. Predictive accuracy as percentage of correct inferences in all 190 paired comparisons of the 20 cities, depicted for the participants and the different strategies. Error bars, where existent, are two standard errors of the mean representing variation between the 42 participants. There are no error bars if the ecological cue values were used because here there is no variation between participants.

What can be seen in Figure 3.2 is firstly that the performance of the participants is well in the range of many of the strategies. The upper benchmark is given by TTB and multiple regression using the ecological cue values both to determine the cue order (TTB) or the cue weights (multiple regression) and to make predictions. Both perform equally well, but suffer as soon as the individual cues are used. But they suffer differently. While for TTB it does not matter whether the cue order is determined on the ecological cue values or on each participant's individual cue values, the performance of multiple regression drops much more strongly when the cue weights were determined on the ecological cues and are then applied to the individual cue weights. Then it performs worse

than all instantiations of TTB using individual cue values. If, however, cue weights are fitted to each individual, multiple regression outperforms TTB using individual cue values. Note, however, that this is a far more complex strategy taking much more information into account. While TTB only uses on average 5.1 pieces of information, multiple regression always looks up all 22 pieces of information and weighs them in a much more complex manner. Minimalist only looks up 3.2 pieces of information on average, and its performance is worse than the performance of TTB and multiple regression. Interestingly, for Minimalist it does not matter whether it relies on ecological or individual cue values.

What do these performances of these strategies indicate for the performance of retrieval-based strategies? First of all, the upper overall benchmark is 85%, which describes the performance of both TTB and multiple regression given the ecological cue values. Since the retrieval-based strategies need to rely on individual cue values, a more realistic benchmark is the performance of the strategies given the individual cue values. These are somewhere between 73.1% (multiple regression with ecological cue weights and individual cue values), 74.5%-75.8% for the two variants of TTB using individual cue values (either with individual or ecological cue orders), and 80.3%, as achieved by multiple regression using cue weights fitted to each individual and individual cue values.

The performance of DifferX. The performance of DifferX and its random counterpart is depicted in Figure 3.3. The X-axis plots the decision threshold, while the Y-axis plots either the predictive accuracy or the frugality. The Y-axis plotting predictive accuracy is limited to 85%, which is the upper benchmark performance of either multiple regression or TTB using ecological cue values.

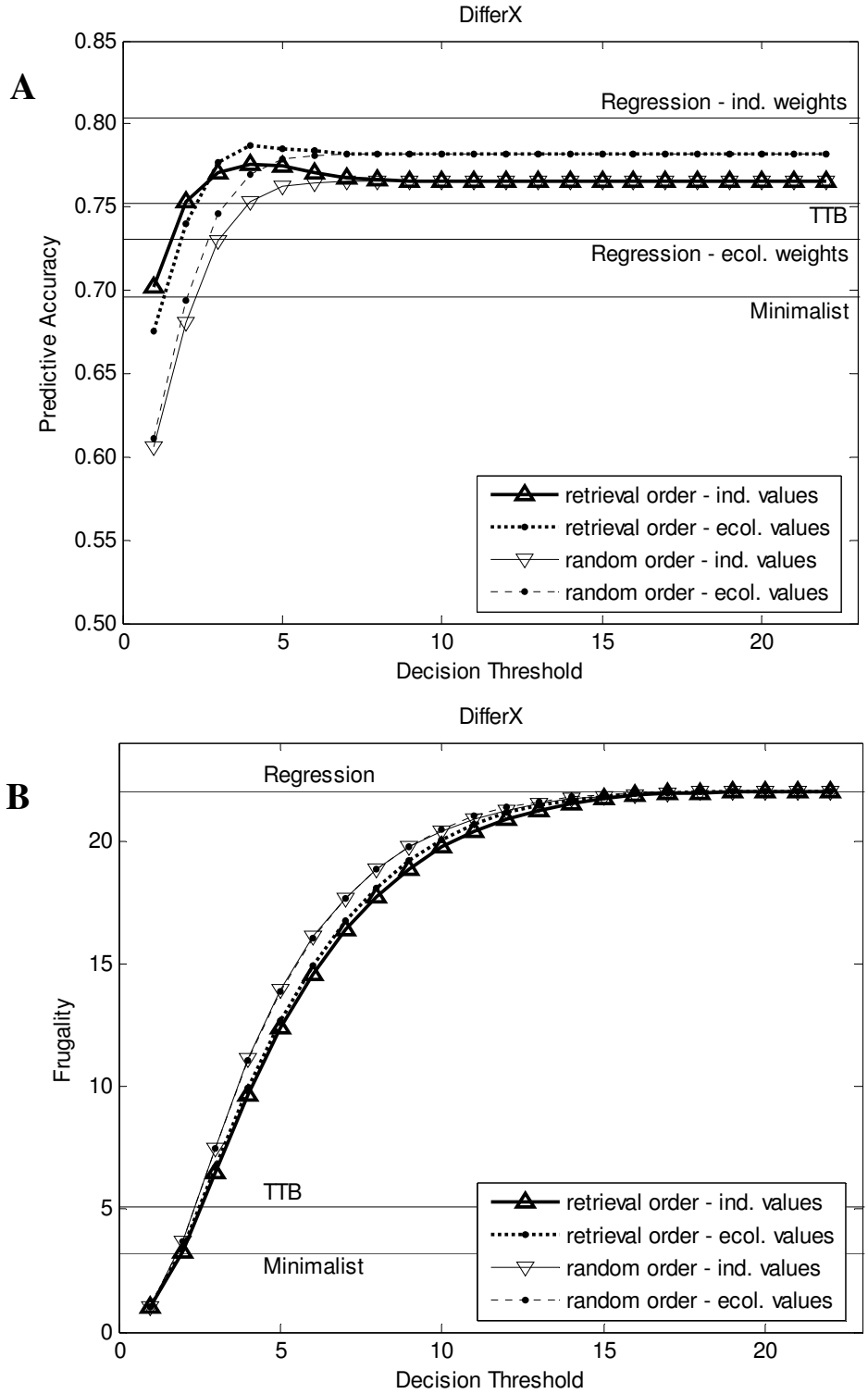

Figure 3.3. Comparison of the retrieval-based strategy DifferX on A) predictive accuracy and B) frugality to their counterpart strategies using a random cue order. The accuracy and frugality of the competitors using individual cue values are indicated by the solid horizontal lines.

The four different curves depicted represent all four combinations of cue order (retrieval-based vs. random) and of cue values (individual vs. ecological). The horizontal solid lines represent the performances of the competitors using individual cue values. The predictive accuracies of multiple regression are further differentiated depending on whether multiple regression uses ecological cue weights or individual cue weights (for frugality this does not make a difference). For TTB, there is no difference between an ecological and an individual cue order, since the predictive accuracies of those two are very close to each other (75.8% vs. 74.5%). Therefore, for TTB, the average of those two predictive accuracies (75.15%) is depicted.

The most striking result that can be observed is that the DifferX strategy using the retrieval order is much better than its equivalent using a random cue order, as long as not too many pieces of information are looked up. The difference is especially pronounced (5% to 10%) when only very few pieces of information are taken into account. Only if at least two thirds of the information is looked up, the random order is as good as the retrieval-based order.

When looking up information in the order of retrieval, looking up one single piece of information is already as successful as Minimalist, which looks up 3.2 pieces of information on average. Furthermore, a Differ2 strategy (i.e., a DifferX strategy with a decision threshold of 2) is as successful as TTB with only looking up as many pieces of information as Minimalist (i.e., approximately two pieces less than TTB) and without any knowledge about cue validities (or about their order).

Congruent with the hypothesis that using the retrieval order might protect people from relying on incorrect cue values (which are retrieved more slowly), the retrieval-based strategy does not benefit much if the ecological instead of the individual cue values are employed. While for the random order strategies using the ecological cue values is already beneficial when only a few pieces of information are considered, DifferX only benefits from the ecological cue values when approximately at least half of the information available is taken into consideration. Thereby, it is much more robust against incorrect cue values as TTB or multiple regression, both of which suffered from a large drop in predictive accuracy if the individual instead of the ecological cue values were used (see Figure 3.2).

These results are similar to the results of other classes of retrieval-based strategies which are reported in Appendix B. They are slightly less successful, in that they cannot be

as accurate as TTB while at the same time being more frugal. All of the strategies are as good as or better than multiple regression with ecological cue weights with a decision threshold which is equal or larger than 2. However, none of the strategies can be as accurate as multiple regression if its weights are fitted to each individual. Nonetheless, they come reasonably close with taking only a fraction of the information into account and without considering the importance of this information.

With regard to frugality, DifferX is the only strategy which is slightly more frugal when relying on a retrieval order rather than a random order. For the other strategies (see Appendix B) this is not the case.

It is somewhat surprising that even DifferX as the most successful retrieval-based strategy falls slightly short of the participants' average performance on the inference task (which was almost 80%). Although the general level of criterion knowledge (i.e., knowing approximately how many inhabitants a city has) was low, it is possible that people have some idea about the order of the cities with regard to city size. For some of the paired comparisons people thus might not have needed to retrieve any cues but could answer more or less directly, which could explain why they outperformed cue-based strategies. For example, people might simply know that Cologne is one of the largest cities in Germany and is thus larger than Trier, without being able to specify an approximate number of inhabitants, but also without having to retrieve any cues. Additionally, they could have also relied on cues that were not assessed by the study.

3.3 General Discussion

In the introduction of the chapter I have speculated that sequential cue-based strategies could exploit the reflection of the environment in their memory system by relying on the fluency of cues to order them. I have hypothesized that there are certain features of these fluencies which such strategies could exploit. First, I expected that positive cue values (i.e., cue values indicating the presence of a cue) should be more fluently available for large objects. For negative cue values (i.e., cue values indicating the absence of a cue), this should be exactly reversed: They should be more fluently available for small objects. Furthermore, I speculated that correct cue knowledge should be more fluently processed than incorrect cue knowledge. This could benefit a strategy that orders the cues by their fluency because it makes such a strategy more robust against incorrect cue knowledge – it will often not consider incorrect cue knowledge at all.

The goal of the Experiment was to assess the fluency of different cues for different objects, which was done using an environment consisting of 20 German cities described by 11 cues. The Experiment assessed response times to questions regarding cue knowledge of the kind "Does object X have cue Y?" The assumption was that the more quickly a person could respond to these questions, the more fluently she or he will have retrieved the cue value for that particular object. The global analyses of the Experiment broadly confirmed the hypotheses about features of fluency that a strategy that orders information by fluency could exploit.

In a second step, the fluencies assessed in the Experiment were used to simulate the performance of such strategies using fluency to order information. I have called those strategies *retrieval-based strategies.* To summarize the results, looking up cues in order of retrieval largely benefits heuristic inferences in the environment of German cities. The retrieval order resulted in better performance than otherwise equivalent strategies that look up cues in random order. Also congruent with the hypotheses, the retrieval-based strategies did not suffer from applying individual (and thus sometimes incorrect) cue values. Comparing the retrieval-based strategies to the other competitors revealed that the retrieval order can fuel strategies to perform as well as TTB without having any knowledge about cue validities. A Differ2 strategy could even match TTB with looking up less information. The comparison with multiple regression is somewhat more difficult. Retrieval-based strategies could easily outperform multiple regression if its weight were determined on the ecological cue values and then applied to the cue values of each individual. However, if the weights were fitted to the cue values of each individual, multiple regression outperformed all retrieval-based strategies by some percentage points.

In sum, the results show that the fluency with which people retrieve cues could potentially be used successfully to order information. What has been often considered to be a bias due to the "availability heuristic" that relies on what comes to mind first or more easily (e.g., Tversky & Kahneman, 1973, 1974) actually proves to be useful if regarded from an ecological perspective. In the introduction to the famous book by Kahneman, Slovic, and Tversky (1982), Tversky and Kahneman (1982) themselves wrote that in "general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors" (p. 3). However, after having said this, their work focused almost exclusively on instances where the heuristics did not work, often leaving their readers with the impression that cognitive illusions and biases are the rule rather than the exception.

In contrast to this view, I believe that it is important to study *when* and *why* (and not *if*) certain heuristic principles can be successful, and when not. To be able to *evaluate* a cognitive process conclusively one therefore needs to consider the match between the cognitive process and the environment in which it operates (Gigerenzer et al., 1999; Simon, 1990). Thus, to evaluate ideas more broadly that have their origin in the availability heuristic and that were precisely defined here as the fluency of cues, I have tried to apply Brunswik's (1956) ideas of representative design in this chapter, which is probably best defined in his following quote: "Generalization of results concerning…the variables involved must remain limited unless the range, but better also the distribution…of each variable, has been made representative of a carefully defined set of conditions" (Brunswik, 1956, p. 53).

Although I investigated a real world environment here, the conclusions and generalizations I drew from the results certainly remain limited. So far, I have only shown that the fluency of cues could successfully fuel strategies in one environment. Naturally, this can be only a first step, and an obvious next step is trying to extend these results to other environments. Furthermore, the question remains open whether people actually do use such retrieval-based strategies. This chapter did not deal with this descriptive question but rather focused on the prescriptive question of whether people, in principle, could successfully use retrieval-based strategies. There were two reasons for that.

First, there exists a huge body of literature following the seminal work by Tversky and Kahneman (1973) demonstrating that what comes to mind first actually has a major impact on people's decisions. Second, and more importantly, it could be argued that the design of the Experiment in this chapter does not really allow one to decide conclusively which strategy someone is using because the set of objects was chosen to represent the real environment, not to differentiate optimally between different strategies.

I believe this to reflect a more general trade-off that often has to be made between representative and controlled design. If one focuses on a representative design that allows for processes to be evaluated, one often loses the ability to describe precisely what people are doing. In the Experiment described in this chapter, I used real objects described on real cues. That is, I studied what people have learnt outside of the laboratory, which is important if one wants to make an ecological argument. In turn, with regard to the question how people actually make inferences, it cannot be ruled out that they use further knowledge not assessed in the study. Furthermore, a representative set of comparisons

between objects as used in the Experiment in this chapter will rarely coincide with a set of comparisons that allows for discriminating between strategies.

In Chapter 2, I used a controlled design which was not necessarily representative of decision-making outside a laboratory setting. Participants learnt everything they needed to answer the questions in the experiments within the laboratory. The comparisons were designed to discriminate well between different strategies. This allowed for showing that many people used simple heuristics in memory-based decisions and that they process information sequentially. But it did not enable an evaluation of whether these heuristics are useful or not.

In sum, I strongly believe that it is important to regard the principles of controlled and representative design to be complementary in nature. One needs controlled design to *understand* and *describe* cognitive processes, and one needs representative design to *evaluate* them. Hopefully, the interplay of Chapters 2 and 3 has contributed both to the understanding and the proper evaluation of cognitive processes in memory-based multi-attribute decisions.

Summary and Conclusion

The starting point of the dissertation was the question how people search for information in memory when they make decisions. Following the perspective of ecological rationality (e.g., Gigerenzer et al., 1999), successful decision strategies are anchored both in the human mind and in the environment. Adaptive decision making thus requires that people adapt their strategies both to the structure of the environment and to the limitations of the cognitive system. In this regard, I am sympathetic to the view put forward for example by Schooler and Hertwig (2005) that these limitations may be functional. Among other functions, they shape how people search for information in memory by facilitating certain ways of searching for information, but hindering others.

In Chapter 1, I have explored the counterintuitive finding that people with a lower short-term memory capacity outperform people with a higher short-term memory capacity in a correlation detection task (Kareev et al., 1997). The task consisted of predicting, on many trials, which of two objects (X or O) an envelope (which was either red or green) contained. There was a correlation between the color of the envelope and the probability of encountering the objects. For example, there were more Xs in red envelopes and more Os in green envelopes. Detecting this correlation thus helps people to improve their predictions.

I have disentangled two potential explanations for this interesting finding of a low capacity advantage, both of which share the idea that people search for information in memory and that this search is constrained by capacity limitations. Kareev et al.'s original explanation, the *small sample hypothesis*, was that low digit spans based their estimates of the correlation on smaller samples from the environment and thereby perceive it as more extreme and detect it earlier. This hypothesis builds on the statistical finding that small samples tend to overestimate correlations.

However, small samples also bear a higher risk of false alarms, making it unclear whether they are really advantageous in this regard (R. B. Anderson et al., 2005; Juslin & Olsson, 2005). Furthermore, correlation estimates have been observed to increase with sample size (e.g., Clément et al., 2002; Shanks, 1985, 1987), which is the opposite of what would be expected by the small sample hypothesis.

Looking for an alternative explanation, I recognized that Kareev et al.'s task was very similar to classical binary choice probability learning tasks which are as simple as it can get: people have to predict one of two events with a different probability of occurring. For example, event E1 could occur with a probability of p(E1) = .75, while event E2 only occurs with p(E2) = 1 – p(E1) = .25. Given that the succession of events is conditionally independent, the best people could do is always predicting the more frequent event E1. This strategy is called *maximizing* and would yield an average accuracy of 75%. However, the modal strategy is *probability matching*, that is, to predict the events in proportion to their probability of occurrence, with an expected accuracy of only 62.5% on average ($.75^2 + .25^2$).

Why do people fail to see the optimal solution in such a simple task? The typical assumption is that people are not smart enough (e.g., West & Stanovich, 2003). In contrast with that view, however, there is convergent evidence showing that beings who have lower cognitive capacities, such as children, pigeons, or people who are distracted by a secondary task, are more likely to behave "rationally" and maximize than the average human adult (e.g., Weir, 1964; Wolford et al., 2004). These results that lower cognitive capacities actually foster maximizing instead of preventing it invited a different view: Probability matching is the result of a more complex strategy – people explore the space of hypotheses how to improve performance on the task. One hypothesis people typically have in those tasks is that there are patterns in the sequence, and any reasonable pattern tends to match the probabilities. Since there are no patterns, however, searching for patterns is counterproductive. Therefore, people who do not search for patterns, for example because of capacity limitations, are more likely to settle on maximizing and will be more successful.

Thus, the low capacity advantage described by Kareev et al. (1997) could be the same kind of phenomenon as the less-is-more effect in probability learning. People with lower cognitive capacities make simpler predictions, which are more successful in this task. I modeled this alternative hypothesis, which I have called *predictive behavior hypothesis*, and Kareev et al.'s original small sample hypothesis of an exaggerated perception of correlations in ACT-R, a cognitive architecture developed by Anderson and colleagues (e.g., J. R. Anderson et al., 2004). The models predicted that simpler predictions impair performance if the environment changes, while a more exaggerated perception of correlation is advantageous to detect a change.

Congruent with differences in the way participants make predictions, two experiments revealed a low capacity advantage before the environment changes, but a high capacity advantage afterwards. Similarly, a third experiment could show that people who maximize more strongly because they are distracted by a secondary task have trouble adapting to a changing environment. The less-is-more effect comes with a price in an unstable environment. Furthermore, it comes with the price that people are more prone to false alarms, which was additionally predicted by the model implementation of the predictive behavior hypothesis and was supported by a fourth experiment.

So probability matching may not be as stupid as it initially appears. First of all, it seems to be a good strategy in situations where the individual is not alone (Gallistel, 1990; Thujisman et al., 1995). Furthermore, the process underlying probability matching, searching for patterns, is usually smart, because often the cost of missing a non-random sequence could well be higher than the price of detecting patterns where there are none (Lopes, 1982). But it looks stupid in stationary binary choice tasks with conditions that rarely hold outside of psychological laboratories and casinos (Ayton & Fischer, 2004). Probability matching, or its underlying process, is smarter than it appears at first glance.

Many decisions we have to face, however, will neither be as simple as the binary choice paradigm nor depend on such a dearth of information. Therefore, Chapter 2 dealt with memory-based decisions in a more complex environment with several cues. The work in this chapter was inspired by the work of Bröder and Schiffer (2003b, 2006) who implemented the idea of memory search in cue based decisions by introducing a cue-learning paradigm. Here, participants had to learn cue values about several objects prior to decisions about which of two objects is larger on a criterion. During the decision phase, no cues were visible, which required participants to retrieve them from memory.

With this paradigm, Bröder and Schiffer (2003b, 2006) provided a method of studying memory-based decisions, which were assumed to be the rule rather the exception by many (e.g., Gigerenzer & Todd, 1999), but were only investigated rarely. Instead, most research focused on inferences from givens where information is provided by the experimenter, for example on the computer screen. Congruent with the findings reported in Chapter 1 that memory demands may increase the prevalence of simple decision strategies, the results from Bröder and Schiffer show that that the need to retrieve cue information from memory induced the use of simple decision strategies, in particular when working memory load was high.

Contrary to studying inferences from givens, however, studying inferences from memory comes with the problem that the decision process is not observable directly. To classify people according to the different strategies they assumedly use, Bröder and Schiffer (2003b, 2006) thus have relied solely on the outcome of the decisions. Therefore, I reanalyzed response times in the data from five experiments conducted by Bröder and Schiffer to investigate whether they convergently supported the outcome-based strategy classification. I also conducted one new experiment to disentangle a potential confound. The idea behind the response time analyses was that the different strategies make different qualitative predictions about patterns of response times. A noncompensatory lexicographic strategy such as Take The Best (TTB) makes the prediction that people process cues in order of their validity and stop as soon as the first cue discriminates between the two objects. Thus, for people who apply TTB, paired comparisons where the most valid cue discriminates should be quicker than paired comparisons where only the 2nd, 3rd, or 4th valid cue discriminate. Compensatory strategies (which include Franklin's rule, FR, and Dawes's rule, DR), in contrast, do not predict an increase in response time that depends on the position of the most valid discriminating cue, since they assume that people always look up all cues for both objects (although also compensatory strategies can make use of validity order; see Rieskamp & Otto, 2006).

The response time patterns nicely fitted the outcome-based strategy classifications: TTB users showed the largest increase in response times depending on how many cues needed to be looked up according to TTB, while there was either no or only a weak increase for users of compensatory strategies. Furthermore, people classified as using the more complex FR were slower than people classified as using the simpler DR, and people who were classified as guessing were fastest. Some people applied a one-reason decision making strategy similar to TTB, but did not apply the validity order. Instead, they processed cues in the order in which this was easiest to accomplish, namely the order in which they were learnt.

In other words, these people relied on the fluency with which they retrieved the cues to order them. In this experiment, the fluency of retrieving information depended only on the learning order, and thereby was not informative. Outside of the laboratory, however, the fluency of retrieving information is informative because it strongly depends on the frequency and recency of encountering information in the environment (J. R. Anderson & Schooler, 1991). In many domains, objects with larger criterion values will

be more often mentioned in the media, which could be for example shown for geographical objects such as cities (Goldstein & Gigerenzer, 2002), but also for political parties (Marewski, Gaissmaier, Dieckmann, Schooler, & Gigerenzer, 2005). Congruently, Schooler and Hertwig (2005) could show that the fluency of recognizing an object (here: a German city) could be used to infer this object's size.

In Chapter 3, I extended the idea that fluency is informative to the level of cues. In particular, I addressed the question whether people could in principle use the fluency with which they retrieve cues to order those cues in cue-based strategies similar to TTB. This is important because simple cue-based heuristics such as TTB owe much of their success to a correct order in which cues are considered (here: by cue validity). This prerequisite of TTB makes it a more difficult strategy than it appears to be at first glance (Juslin & Persson, 2002).

I conducted an experiment to assess the accuracy of people's cue knowledge concerning real world objects (here: German cities), and to assess how fluently they retrieve the cues. People were asked questions about attributes of German cities, all of which had the format "Does city X have cue Y?" The assumption was that the faster people will be able to answer this question, the more fluently available this cue value is to them.

The results of this experiment revealed that the fluency with which cues about objects come to mind is indeed informative. People retrieved positive cue values (i.e., cue values indicating the presence of a cue, such as, e.g., an airport) more quickly for larger cities. This is presumably the case because they have encountered more pieces of information about larger cities in the environment. For negative cue values (i.e., cue values indicating the absence of a cue), this was exactly the other way around: Negative cue values were more fluently available for smaller cities. Furthermore, retrieving incorrect cue values was slower on average, which is congruent with findings in the memory literature that giving incorrect answers often takes longer (e.g., Ratcliff & Smith, 2004).

All of these findings could be exploited by strategies using the fluency of cues to order them (which I have called *retrieval-based strategies*): If a person is asked to infer which of two cities is larger, this person will, on average, more quickly come up with positive cues favoring the city which is actually larger. Moreover, this person will, on average, more quickly come up with negative cues speaking against the city which is

actually smaller. Furthermore, ordering cues by fluency will move incorrect cue values backwards in the order. This could additionally fuel the success of retrieval-based strategies because it prevents those strategies from considering incorrect cue values.

In a next step, I have used these cue fluencies to simulate how well retrieval-based strategies would perform in an inference task. The inference task consisted of all possible paired comparisons between the cities about which cue knowledge was assessed in the experiment. The basic principle behind all of the retrieval-based strategies is that they accumulate evidence for or against the cities in order of cue fluency (starting with the most fluent one) until a decision threshold is reached (e.g., that one city is favored by three pieces of information). These retrieval-based strategies were compared to strategies that are structurally identical except that they looked up cue values in a random order. They were also compared to competitor models such as TTB and multiple regression.

The results of the simulations showed that the retrieval order largely benefited the decision strategies. All of the retrieval-based strategies were much more accurate than their counterparts using a random order, at least as long as less than two thirds of all available information was considered. The difference was most strongly pronounced for low decision thresholds and thereby very frugal strategies. The comparison to the competitor models showed that the retrieval-based strategies can be as successful as TTB or even outperform TTB; one of the retrieval-based strategies (DifferX) could achieve the same accuracy as TTB while at the same time being more frugal. Moreover, all of the retrieval-based strategies could outperform one version of multiple regression in which some kind of crossvalidation was applied (for the details, see section 3.2.3); and while only taking a fraction of the information into account, the retrieval-based strategies even came close to another version of multiple regression in which the weights were extensively fitted to each individual separately.

These results demonstrate that people do not need to know how important different cues are to order them successfully. They can let the environment do the work and rely on how the environment is reflected in their memory by simply considering cues in the order of fluency with which they are retrieved. I believe that the retrieval-based strategies discussed in Chapter 3 are thereby good examples of decision strategies that are both anchored in the environment and in the human mind.

More generally, I hope that my dissertation has contributed to show that humans do neither need complete information nor unlimited time to make good judgments. The

strategies they use are well adapted both to the environment and to the human mind and can thus be successful and simple at the same time. The memory system can help by shaping the way people search for information they have stored. It can guide the search towards useful information and can prevent people from searching too much information, which could be unnecessary or even detrimental.

In the view of William James (1890) the limitations of memory are thereby one important instance of the mind's selectivity, and since he already formulated this thought nicely more than 100 years ago, he shall have the final words of my dissertation:

> This peculiar mixture of forgetting with our remembering is but one instance of our mind's selective activity. Selection is the very keel on which our mental ship is built. And in this case of memory its utility is obvious. If we remembered everything, we should on most occasions be as ill off as if we remembered nothing. It would take as long for us to recall a space of time as it took the original time to elapse, and we should never get ahead with our thinking. (James, 1890, p. 680)

References

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation in humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91*, 112–149.

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

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

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*, 1036–1060.

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, R. B., Doherty, M. E., Berg, N. D., & Friedrich, J. C. (2005). Sample size and the detection of correlation: A signal detection account. *Psychological Review, 112*, 268–279.

Arrow, K. J. (1958). Utilities, attitudes, choices: A review note. *Econometrica, 26*, 1–23.

Ayton, P., & Fischer, I. (2004). The hot-hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition, 32*, 1369–1378.

Bauer, M. (1972). Relations between prediction- and estimation-responses in cue-probability learning and transfer. *Scandinavian Journal of Psychology, 13*, 198–207.

Bergert, F. B., & Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-the-best model of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 107–129.

Betsch, T., Haberstroh, S., Glöckner, A., Haar, T., & Fiedler, K. (2001). The effects of routine strength on adaptation and information search in recurrent decision making. *Organizational Behavior and Human Decision Processes, 84*, 23–53.

Bitterman, M. E., Wodinsky, J., & Candland, D. K. (1958). Some comparative psychology. *American Journal of Psychology, 71*, 94–110.

Brackbill, N., & A. Bravos (1962). Supplementary report: The utility of correctly predicting infrequent events. *Journal of Experimental Psychology, 62*, 648–649.

Bröder, A. (2000). Assessing the empirical validity of the "Take-The-Best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1332–1346.

Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology:Learning Memory, and Cognition, 29*, 611–625.

Bröder, A., & Schiffer, S. (2003a). Bayesian strategy assessment in multi-attribute decision making. *Journal of Behavioral Decision Making, 16*, 193–213.

Bröder, A., & Schiffer, S. (2003b). Take the best versus simultaneous feature matching: Probabilistic inferences from memory and effects of representation format. *Journal of Experimental Psychology: General, 132*, 277–293.

Bröder, A., & Schiffer, S. (2006). Stimulus format and working memory in fast and frugal strategy selection. *Journal of Behavioral Decision Making, 19*, 361-380.

Brunswik, E. (1939). Probability as a determiner of rat behavior. *Journal of Experimental Psychology, 25*, 175–197.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments.* Berkeley: University of California Press.

Bundesliga (2006). *Clubs und Spieler.* Retrieved from http://www.bundesliga.de/de/liga/clubs/index.php on April 3rd, 2006.

City Population (2006). *Deutschland.* Retrieved from http://www.citypopulation.de/Deutschland_d.html on April 4th, 2006.

Clément, M., Mercier, P., & Pasto, L. (2002). Sample size, confidence, and contingency judgement. *Canadian Journal of Experimental Psychology, 56*, 128–137.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review, 12*, 769–786.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-185.

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

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582.

De Houwer, J., & Beckers, T. (2002). A review of recent developments in research and theory on human contingency learning research. *Quarterly Journal of Experimental Psychology, 55B*, 289–310.

Deese, J. (1960). Frequency of usage and number of words in free recall: The role of association. *Psychological Reports, 7*, 337–344.

Derks, P. L., & Paclisanu, M. I. (1967). Simple strategies in binary predictions by children and adults. *Journal of Experimental Psychology, 73*, 278–285.

Deutsche Bahn (2006). *Streckenkarten Fernverkehr.* Retrieved from http://www.bahn.de/-S:PtVOSN:ek54nNNNP-5EJdNNNOEM/p/view/planen/reiseplanung/streckenkarten_fernv.shtml on March 20th, 2006.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*, 959 – 988.

Dieckmann, A., & Todd, P. (2004). Simple ways to construct search orders. *Proceedings of the 26th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Dougherty, M. R. P., & Hunter, J. E. (2003). Hypothesis generation, probability judgment and individual differences in working memory capacity. *Acta Psychologica, 113*, 263–282.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106*, 180 - 209.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General, 128*, 309–331.

Erev, I. (1998). Signal detection by human observers: A cutoff reinforcement learning model of categorization decisions under uncertainty. *Psychological Review, 105*, 280–298.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112*, 912–932.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83*, 37–64.

Estes, W. K., & Straughan, J. H. (1954). Analysis of a verbal conditioning situation in terms of statistical learning theory. *Journal of Experimental Psychology, 47*, 225–234.

Fantino, E., & Esfandiari, A. (2002). Probability matching: Encouraging optimal responding in humans. *Canadian Journal of Experimental Psychology, 56*, 58–63.

Fiorina, M. P.(1971). A Note on Probability Matching and Rational Choice. *Behavioral Science 16*, 158-166.

Gallistel, C. R. (1990). *The Organization of Learning.* Cambridge, MA: MIT Press.

Geoff, W., Woodward, G., Stevens, A., & Stinson, C. (2003). Using overt rehearsals to explain word frequency effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 186–210.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review, 103*, 592-596.

Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. In D. Koehler & N. Harvey (Eds.), *Handbook of judgement and decision making* (pp. 62–88). Oxford: Blackwell.

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

Gigerenzer, G., & Selten, R. (2001). Rethinking rationality. In G. Gigerenzer & R. Selten, (Eds.), *Bounded rationality. The adaptive toolbox* (pp. 1–12). Cambridge, MA: The MIT Press.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: the adaptive toolbox. In G. Gigerenzer, P. M. Todd & the ABC Research Group, *Simple heuristics that make us smart* (pp. 3–34). New York: Oxford University Press.

Gigerenzer, G., & Todd, P. M. and 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.

Gonzalez, C., Lerch, F. J., & Lebiere, C. (2003). Instance-based learning in real-time dynamic decision making. *Cognitive Science, 27*, 591–635.

Goodnow, J. J. (1955). Determinants of choice-distribution in two-choice situations. *American Journal of Psychology, 68*, 106–116.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple-read out model. *Psychological Review, 103*, 518–565.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice.* Oxford, UK: Oxford University Press.

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 344–354.

Herlitz, A., Nilsson, L.-G., & Bäckman, L. (1997). Gender differences in episodic memory. *Memory & Cognition, 25*, 801–811.

Herrnstein, R. J., & Loveland, D. H. (1975). Maximizing and matching on concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior, 24*, 107–116.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*, 383–403.

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, UK: Wiley.

Hinson, J. M., & Staddon, J. E. R. (1983). Matching, maximizing and hillclimbing. *Journal of the Experimental Analysis of Behavior, 40*, 321–331.

Hochschulkompass (2006). *Hochschule suchen.* Retrieved from http://www.hochschulkompass.de/kompass/xml/index_hochschule.htm on April 6th, 2006.

Humphreys, L. G. (1939). Acquisition and extinction of verbal expectations in a situation analogous to conditioning. *Journal of Experimental Psychology, 25*, 294–301.

Hyman, R & Jenkin, N. S. (1956). Involvement and set as determinants of behavioral stereotypy. *Psychological Reports, 2*, 131-146.

Irwin, J. R., & McClelland, G. H. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research, 40*, 366–371.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition – Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1311-1334.

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

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

Jarvik, M. E. (1951). Probability learning and a negative recency effect in the serial aniticpation of alternative symbols. *Journal of Experimental Psychology, 41*, 291–297.

Johnson, E. J., & Payne, J. W. (1985). Effort and accuracy in choice. *Management Science, 31*, 395–414.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard University Press.

Juslin, P., & Olsson, H. (2005). Capacity limitations and the detection of correlation: Comment on Kareev (2000). *Psychological Review, 112*, 256–267.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26*, 563–607.

Juslin, P., Olsson, H., & Olsson, A. C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, 132, 133–156.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 336–358.

Kareev, Y. (1995a). Positive bias in the perception of covariation. *Psychological Review, 102*, 490–502.

Kareev, Y. (1995b). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition, 56*, 263–269.

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

Kareev, Y. (2004). On the perception of consistency. Psychology of Learning and Motivation: *Advances in Research and Theory, 44*, 261–285.

Kareev, Y. (2005). And yet the small-sample effect does hold: Reply to Juslin and Olsson (2005) and Anderson, Doherty, Berg, and Friedrich (2005). *Psychological Review, 112*, 280–285.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General, 126*, 278–287.

Kirkpatrick, S., Gelatt Jr., C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science, 220*, 671–680.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity?! *Intelligence, 14*, 389–433.

Leibniz, G. W. (1951) Toward a universal characteristic. In P. P. Wiener (ed.), *Leibniz: Selections*. Scribner's Sons. (Original work published in 1677.)

Lewin, C., Wolgers, G., & Herlitz, A. (2001). Sex differences favoring women in verbal and non-verbal, but not in visuospatial episodic memory. *Neuropsychology, 15*, 165–173.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492–527.

Lopes, L. L. (1982). Doing the impossible: A note on induction and the experience of randomness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 626–636.

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

Marewski, J. N., Gaissmaier, W., Dieckmann, A., Schooler, L. J., & Gigerenzer, G. (2005, August). *Ignorance-based reasoning? Applying the recognition heuristic to*

elections. Paper presented at the 20th Biennial Conference on Subjective Probability, Utility and Decision Making, Stockholm.

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

Maule, A. J. (1994). A componential investigation of the relation between structural modelling and cognitive accounts of human judgement. *Acta Psychologica, 87*, 199–216.

Messen.de (2006). *Messe Deutschland Suche.* Retrieved from http://www.messen.de/messe-suche-deutschland.php on April 2nd 2006.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Mitchell, T. R., & Beach, L. R. (1990). " . . .Do I love thee? Let me count . . .": Toward an understanding of intuitive and automatic decision making. *Organizational Behavior and Human Decision Processes, 47*, 1–20.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General, 130*, 621–640.

Morse, E. B., & Runquist, W. N. (1960). Probability-matching with an unscheduled random sequence. *The American Journal of Psychology, 73*, 603-607.

MSN Encarta (2006). *Bundesrepublik Deutschland. 6.4. Industrie.* Retrieved from http://de.encarta.msn.com/encyclopedia_761576917_7/ Bundesrepublik Deutschland.html on March 20th, 2006.

Mutter, S. A., & Williams, T. W. (2004). Aging and the detection of contingency in causal learning. *Psychology and Aging, 19*, 13–26.

Myers, J. L. (1976). Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Approaches to human learning and motivation* (pp. 171–205). Hillsdale, NJ: Erlbaum.

Neimark, E. D., & Shuford, E. H. (1959). Comparison of predictions and estimations in a probability learning situation. *Journal of Experimental Psychology, 57*, 294–298.

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 53–65.

Newell, B. R., Rakow, T., Weston, N. J., & Shanks, D. R. (2004). Search strategies in decision-making: The success of „success". *Journal of Behavioral Decision Making, 17*, 117–137.

Newell, B. R., Weston, N. J., & Shanks, D. R. (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone "takes-the-best". *Organizational Behavior and Human Decision Processes, 91*, 82–96.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266–300.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 211–233.

Overman, W. H. (2004). Sex differences in early childhood, adolescence and adulthood on cognitive tasks that rely on orbital prefrontal cortex. *Brain and Cognition, 55*, 134–147.

Pachur, T., Bröder, A., & Marewski, J. N. (2006, March). *Ignorieren wir Informationen bei Ignoranz-basierten Inferenzen?* Paper presented at the 48th Tagung experimentell arbeitender Psychologen, Mainz, Germany.

Parr, W., & Mercier, P. (1998). Adult age differences in on-line contingency judgments. *Canadian Journal of Experimental Psychology, 52*, 147–158.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 534–552.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* Cambridge: Cambridge University Press.

Pinker, S. (1997). *How the Mind Works.* New York: Norton.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59-108.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two choice reaction time. *Psychological Review, 111*, 333-367.

Reavis, R., & Overman, W. H. (2001). Adult sex differences on a decision–making task previously shown to depend on the orbital prefrontal cortex. *Behavioral Neuroscience, 115*, 196–206.

Rieskamp, J., & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell? In G. Gigerenzer, P. M. Todd & the ABC Research Group, *Simple heuristics that make us smart* (pp. 141–167). New York: Oxford University Press.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135*, 207-236.

Rieskamp, J., Busemeyer, J.R., Laine, T. (2003). How do people learn to allocate resources? Comparing two learning theories. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29*, 1066–1081.

Sanford, A. J., & Maule, A. J. (1973). The concept of general experience: Age and strategies in guessing future events. *Journal of Gerontology, 28*, 81–88.

Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology, 32*, 219–250.

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112*, 610–628.

Serwe, S., & Frings, C. (2006). Who will win Wimbledon? The recognition heuristic in predicting sports events. *Journal of Behavioral Decision Making , 19*, 321–332.

Shaklee, H., & Mims, M. (1982). Sources of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 8*, 208–224.

Shanks, D. R. (1985). Continuous monitoring of human contingency judgement across trials. *Memory & Cognition, 13*, 158–167.

Shanks, D. R. (1987). Acquisition functions in causality judgement. *Learning and Motivation, 18*, 147–166.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15*, 233–250.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior, 6*, 156 –163.

Siegel, S., & Goldstein, D. A. J. (1959). Decision-making behavior in a two-choice uncertain outcome situation. *Journal of Experimental Psychology: General, 57*, 37-42.

Simon, H. A. (1990). Invariants of human behavior. Annual Review of Psychology, 41, 1–19.

Singer, E. (1967). Ability and the use of optimal strategy in decisions. *American Journal of Psychology, 80*, 243–249.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology, 25*, 207–222.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*, 161–88.

Statistisches Bundesamt (2004). *Fachserie 8 / Reihe 5. Verkehr. Seeschifffahrt.* Received electronically on April 24th 2006 from a representative of the "Statistisches Bundesamt".

Statistisches Bundesamt (2005). *Binnenschifffahrt 2004. Umschlagstruktur der wichtigsten Häfen.* Received electronically on April 24th 2006 from a representative of the "Statistisches Bundesamt".

Statistisches Bundesamt (2006). *Fachserie 8 / Reihe 6. Verkehr. Luftverkehr.* Retrieved from https://www-ec.destatis.de/csp/shop/sfg/bpm.html.cms.cBroker.cls?cmspath=home on April, 21st 2006.

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition, 86*, 123–155.

Taatgen, N. A., Lebiere, C., & Anderson, J. R. (2006). Modeling paradigms in ACT-R. In R. Sun (ed.), *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation* (pp. 29-52). Cambridge University Press.

Thuijsman, F. Peleg, B. Amitai, M., & Shmida, A. (1995). Automata, matching and foraging behavior of bees. *Journal of Theoretical Biology, 175*, 305-316.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Tversky, A., & Kahneman, D. (1982). Judgment under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 3–20). Cambridge, UK: Cambridge University Press.

Urban Audit (2006a). *Data that can be accessed.* Retreived from http://www.urbanaudit.org/DataAccessed.aspx on April 4th, 2006.

Urban Audit (2006b).*City selection.* Retrieved from http://www.urbanaudit.org/help.aspx on April 4th, 2006.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior.* Princeton, NJ: Princeton University Press.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14*, 101–118.

Weir, M. W. (1964). Developmental changes in problem-solving strategies. *Psychological Review, 71*, 473–490.

West, R. F., & Stanovich, K. E. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition, 31*, 243–251.

Wilson, W. A., Jr., & Rollin, A. R. (1959). Two-choice behavior of rhesus monkeys in a noncontingent situation. *Journal of Experimental Psychology, 58*, 174–180.

Wolford, G., Miller, M. B., & Gazzaniga, M. (2000). The left hemisphere's role in hypothesis formation. *The Journal of Neuroscience, 20* (RC64), 1–4.

Wolford, G., Newman, S., Miller, M. B., & Wig, G. (2004). Searching for patterns in random sequences. *Canadian Journal of Experimental Psychology, 58*, 221–228.

Yechiam. E., and Busemeyer, J.R. (2005). Comparison of basic assumptions embedded in learning models for experience based decision-making. *Psychonomic Bulletin and Review, 12*, 387–402.

Yellott, J. I., Jr. (1969). Probability learning with noncontingent success. *Journal of Mathematical Psychology, 6,* 541–575.

Appendix

Appendix A

Table A. Definitions of the cues.

Cue	Description and Reference
National / State Capital	The cue "Capital" was divided into "National Capital" (i.e., Capital of Germany) and "Federal State Capital" (i.e., Capital of one of the 16 federal States within Germany).
Expositions	On an exposition database on the internet (Messen.de, 2006) I searched for international expositions taking place in 2006 and 2007. All cities in which at least one such international exposition took place during that period were counted as having international expositions.
Airport	This cue assesses whether a city has an airport or not. Since a city might have a very minor (e.g., sports) airport, only airports were considered about which the "Statistisches Bundesamt" collects data in monthly brochures (Statistisches Bundesamt, 2006). In answer to an email request, a representative of the "Statistisches Bundesamt" clarified the criteria for including airports in the brochures: In 2006, the brochures included those 25 German airports with at least 150 000 passengers per year, a threshold defined by the EU. These airports represent almost 100% of all (international) passenger traffic in Germany.
Stations	I focused on major train stations; that is, train stations where ICE trains (fast long distance trains) stop (Deutsche Bahn, 2006).
Soccer	This cue assesses whether a city had a soccer team in the "Bundesliga" (premier league) in the season 2005/2006 (Bundesliga, 2006).

Industry	The defining criterion of whether a city is an important industrial city was whether a city was mentioned in the MSN-Encarta (2006) in the subsection about industry in Germany.
University	The "Higher education compass" on the web contains a university search function (Hochschulkompass, 2006). For each of the 35 German cities from the Urban Audit database, I looked up whether the city had a university or not. Note that only universities were considered, not arts or music schools, and also "Fachhochschulen" (universities of applied science) did not count.
Harbor	This cue assesses whether a city has a harbor or not. Similar to the airport cue, only harbors were considered that the "Statistisches Bundesamt" considers being "the most important harbors" in Germany for inland water transport and for ocean shipping (Statistisches Bundesamt, 2004, 2005).
Infrastructure	As an indicator for the quality of the infrastructure I used the Urban Audit indicator "Multimodal Accessibility", which assesses how accessible a city is in different ways such as by air, by train, and by car (Urban Audit, 2006a). For the purpose of the study, this continuous indicator was dichotomized at the median multimodal accessibility of all the 35 German cities for which data were available on Urban Audit.
Tourism	As an indicator for the amount of tourism in each city, I used the Urban Audit indicator "tourist overnight stays per resident population" (Urban Audit, 2006a). This indicator divides the number of tourist overnight stays by the number of inhabitants. That is, it assesses whether a city is visited by many tourists relative to its size. For the purpose of the study, this continuous indicator was dichotomized at the median multimodal accessibility of all the 35 German cities for which data were available on Urban Audit.

Appendix

Table B. German city environment.

	Cue	Nat. Capital	Expositions	Airport	Stations	Soccer	Industry	University	Harbor	Infrastructure	St. Capital	Tourism
	Validity	1.00	0.91	0.90	0.89	0.84	0.84	0.75	0.71	0.66	0.64	0.61
	Discr. Rate	0.10	0.52	0.52	0.10	0.51	0.51	0.19	0.52	0.51	0.48	0.51
City	Population											
Berlin	3388477	1	1	1	1	1	1	1	1	1	1	1
Munich	1247873	0	1	1	1	1	1	1	0	0	1	1
Cologne	965954	0	1	1	1	1	1	1	1	1	0	1
Frankfurt M.	643432	0	1	1	1	1	1	1	1	1	0	1
Essen	589499	0	1	0	1	0	0	1	1	1	0	0
Bremen	544853	0	1	1	1	1	1	1	1	0	1	0
Leipzig	497531	0	1	1	1	0	1	1	0	0	0	1
Nuremberg	493553	0	1	1	1	1	0	1	1	0	0	1
Dresden	483632	0	0	1	1	0	0	1	0	0	1	1
Wuppertal	362137	0	0	0	1	0	0	1	0	1	0	0
Bielefeld	328452	0	0	0	1	1	0	1	0	0	0	0
Wiesbaden	271995	0	1	0	1	0	1	0	1	1	1	1
Augsburg	259217	0	1	0	1	0	0	1	0	0	0	0
Freiburg	212495	0	1	0	1	0	0	1	0	0	0	1
Erfurt	201645	0	0	1	1	0	0	1	0	0	1	0
Mainz	185532	0	0	0	1	1	1	1	1	1	1	1
Göttingen	122883	0	0	0	1	0	0	1	0	0	0	0
Moers	107903	0	0	0	0	0	0	0	0	1	0	0
Trier	100180	0	0	0	1	0	0	1	1	0	0	1
Weimar	64409	0	0	0	1	0	0	1	0	0	0	1

Appendix B

In the following, I will describe two other variants of retrieval-based strategies that are similar to the DifferX strategy described in Chapter 3. Instead of requiring a fixed information difference between the two objects, these strategies either require a fixed amount of total information or a fixed amount of information favoring one object. Note that these two strategies and the DifferX strategy reported in Chapter 3 are identical for a decision threshold of 1.

As with the DifferX strategy, these strategies look up information in order of retrieval. Negative information about one object counts as information favoring the other object. That is, negative information about object B favors object A identically, as does positive information about object A.

TallyX – Fixed total information. These strategies are tallying strategies comparable to Dawes's rule (without the assumption of cue-wise comparisons). They assume that people retrieve a fixed amount of total information, that is, a fixed number of pieces of information. As soon as this total information threshold is reached, the strategy simply counts which object is favored by more pieces of information and infers this object to be larger. If there is a tie, the strategy guesses. These strategies will be called *TallyX* in the following, with X specifying the decision threshold (here: the total number of pieces of information the strategy considers before it stops to search for further information). For example, Tally3 would look up three pieces of information and then decide for the option favored by more pieces of information.

TakeX – Fixed information favoring one object. These strategies are evidence accumulation strategies that stop searching for information as soon as one object is favored by enough information, independent of how much information favors the other object. These strategies will be called *TakeX* subsequently, with X again specifying the decision threshold (here: number of information pieces that have to favor one object before the search for information stops). For example, Take3 would search for information until one object is favored by three pieces of information, independent of whether there are zero, one, or two pieces of information favoring the other object. Should all pieces of information be looked up before the decision threshold is reached, the strategy settles for the object that is favored by more pieces of information. Should there be a tie, the strategy guesses.

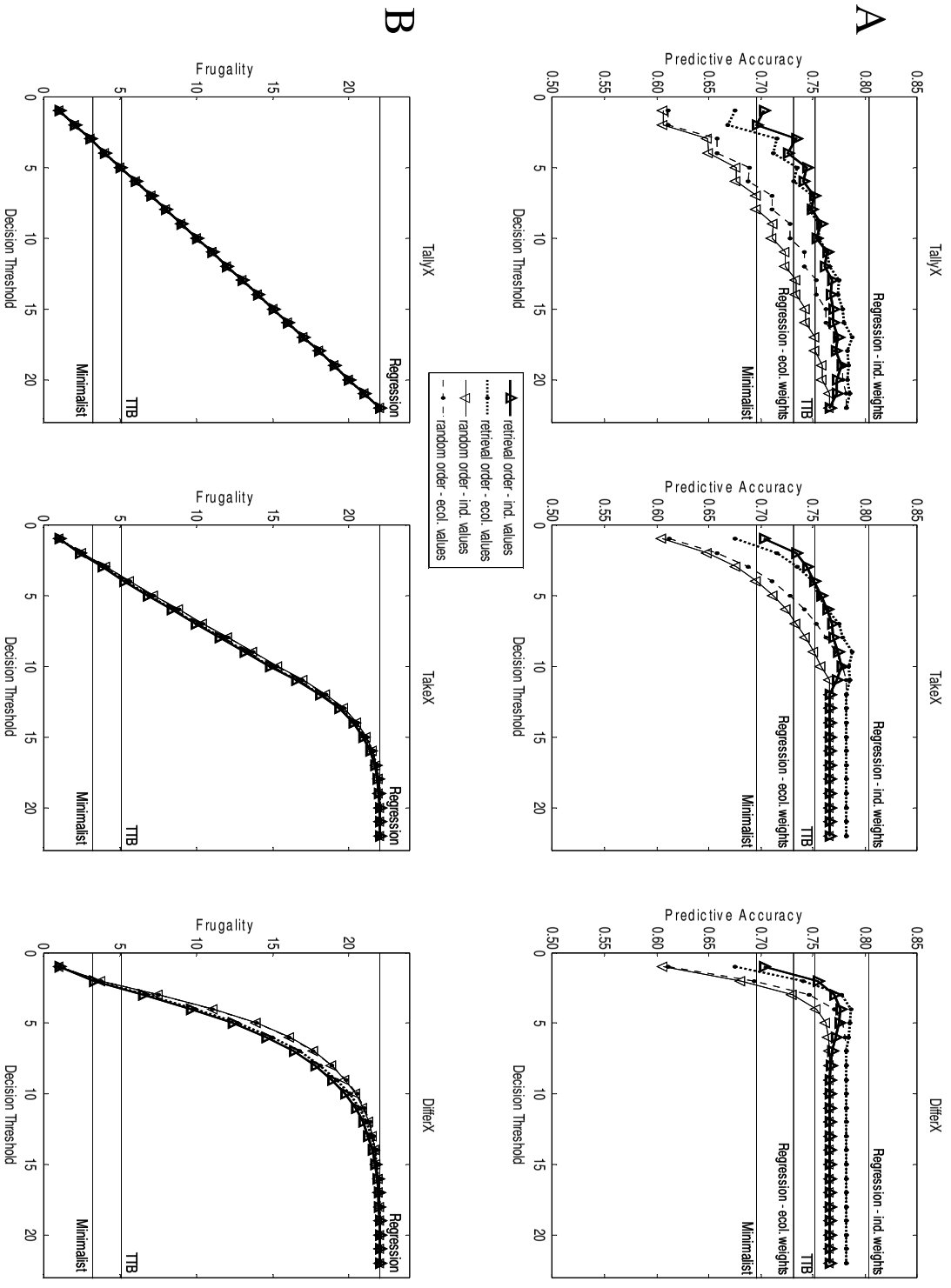

Figure A. Comparison of the retrieval-based strategies on A) predictive accuracy and B) frugality to their counterpart strategies using a random cue order. The accuracy and frugality of the competitors using individual cue values are indicated by the solid horizontal lines.

Deutsche Zusammenfassung (German Summary)

Der Ausgangspunkt der Dissertation war die Frage, wie Menschen nach Informationen im Gedächtnis suchen, wenn sie Entscheidungen treffen. Nach der Perspektive der ökologischen Rationalität (z.B. Gigerenzer et al., 1999) sind erfolgreiche Entscheidungsstrategien sowohl im menschlichen Geist als auch in der Struktur der Umwelt verankert. Adaptives Entscheiden erfordert folglich, dass Menschen ihre Strategien der Struktur der Umwelt und den Beschränkungen des kognitiven Systems anpassen. In dieser Hinsicht teile ich die Ansicht, die zum Beispiel von Schooler und Hertwig (2005) vertreten wird, dass diese Beschränkungen funktionell sein können. Neben anderen Funktionen formen sie, wie Menschen nach Informationen im Gedächtnis suchen, indem sie bestimmte Arten des Suchens nach Informationen erleichtern, aber andere erschweren.

In Kapitel 1 bin ich dem kontraintuitiven Befund nachgegangen, dass Menschen mit einer niedrigeren Kapazität des Kurzzeitgedächtnisses besser bei einer Korrelationsentdeckungsaufgabe abschneiden als Menschen mit einer höheren Kapazität (Kareev et al., 1997). Die Aufgabe bestand darin, in vielen Durchgängen jeweils vorherzusagen, welches von zwei Objekten (X oder O) sich in einem Briefumschlag, der entweder rot oder grün war, befindet. Hierbei bestand eine Korrelation zwischen der Farbe des Umschlags und der Auftretenswahrscheinlichkeit der Objekte, beispielsweise war X häufiger in roten Umschlägen und O häufiger in grünen Umschlägen. Diese Korrelation zu entdecken war folglich hilfreich, um die Vorhersagen zu verbessern.

Ich habe zwei mögliche Erklärungen für diesen interessanten weniger-ist-mehr Befund gegeneinander getestet, die beide auf der Idee basieren, dass Menschen bei dieser Aufgabe nach Informationen im Gedächtnis suchen und dass diese Suche durch Beschränkungen der Gedächtniskapazität begrenzt wird. Kareev et al.' s ursprüngliche Erklärung, die Stichproben-Hypothese, bestand darin, dass Personen mit niedrigerer Gedächtniskapazität ihre Schätzungen der Korrelation auf kleineren Stichproben der Umwelt gründeten. Statistisch gesehen neigen kleine Stichproben dazu, Korrelationen zu überschätzen. Daher, so wurde angenommen, sollten Personen mit niedrigerer Gedächtniskapazität Korrelationen als extremer wahrnehmen und sie früher erkennen. Jedoch bergen kleine Stichproben auch ein erhöhtes Risiko eines falschen Alarms, wodurch unklar ist, ob sie in diesem Zusammenhang wirklich vorteilhaft sind (R.B.

Anderson et al., 2005; Juslin & Olsson, 2005). Außerdem gibt es Untersuchungen, die zeigen, dass die Einschätzung von Korrelationen eher mit einer zunehmenden Zahl an Beobachtungen ansteigen (z.B., Clément et al., 2002; Shanks, 1985, 1987), was der Stichproben-Hypothese exakt entgegensteht.

Auf der Suche nach einer alternativen Erklärung stellte ich fest, dass Kareev et al.'s Aufgabe klassischen binären Wahlaufgaben sehr ähnlich war, die folgendermaßen aussehen: Probanden müssen Durchgang für Durchgang tippen, welches von zwei Ereignissen E1 oder E2 als nächstes auftreten wird. Beide Ereignisse haben hierbei üblicherweise eine unterschiedliche Auftretenswahrscheinlichkeit. Zum Beispiel könnte Ereignis E1 mit einer Wahrscheinlichkeit von p (E1) = .75 auftreten, während Ereignis E2 nur mit p(E2) = 1 - p(E1) = .25 auftritt. Angenommen, die Ereignisreihenfolge sei zufällig. Dann ist das Beste, was Menschen tun können, immer das häufigere Ereignis E1 vorherzusagen. Diese Strategie wird Maximierung genannt und würde eine durchschnittliche Genauigkeit von 75% erbringen. Jedoch ist die am häufigsten beobachtete Strategie, dass die Ereignisse im Verhältnis zu ihrer Auftretenswahrscheinlichkeit vorhergesagt werden. Diese Strategie wird als „probability matching“ bezeichnet und hat nur eine erwartete Genauigkeit von 62.5% ($.75^2 + .25^2$).

Warum schaffen es Menschen nicht, die optimale Lösung in einer solch einfachen Aufgabe zu finden? Die typische Annahme ist, dass Menschen nicht intelligent genug sind (z.B., West & Stanovich, 2003). Im Gegensatz zu dieser Ansicht gibt es jedoch konvergente Evidenz, die zeigt, dass eine niedrigere Gedächtniskapazität bei dieser Aufgabe zu „rationalerem“ Verhalten führt: So schneiden Kinder, Tauben oder Menschen, die durch eine Zweitaufgabe abgelenkt werden besser ab als der durchschnittliche menschliche Erwachsene (z.B., Weir, 1964; Wolford et al., 2004). Diese Resultate führten zu einer anderen Erklärung von „probability matching“: Es scheint nicht das Ergebnis von mangelnder Intelligenz zu sein, sondern das Resultat einer komplizierteren Strategie, nämlich der Exploration von Hypothesen, wie sich die Leistung bei der Aufgabe verbessern ließe. Eine typische Hypothese, die Menschen in diesem Zusammenhang haben, ist die Annahme, dass es Muster in der Abfolge der Ereignisse gibt. Jedes einigermaßen plausible Muster muss beide Ereignisse proportional zu deren Auftretenswahrscheinlichkeiten beinhalten. Da es jedoch in Wirklichkeit keine Muster in der Abfolge gibt, ist die Suche nach Mustern kontraproduktiv. Folglich ist es für

Menschen, die nicht nach Mustern suchen, z.B. aufgrund von Kapazitätsbeschränkungen, wahrscheinlicher, die Maximierungsstrategie anzuwenden und erfolgreicher zu sein.

Der Vorteil einer geringen Kurzzeitgedächtniskapazität bei der Korrelationsentdeckungsaufgabe (Kareev et al., 1997) könnte daher ein ähnliches Phänomen sein wie der weniger-ist-mehr Effekt bei binären Wahlaufgaben. Menschen mit niedrigerer kognitiver Kapazität treffen einfachere Vorhersagen, die in dieser Aufgabe erfolgreicher sind. Ich modellierte diese alternative Hypothese der einfacheren Vorhersagen und Kareev et al.'s ursprüngliche Hypothese der extremeren Wahrnehmung von Korrelationen in ACT-R, einer kognitiven Architektur von Anderson und Kollegen (z.B., J. R. Anderson et al., 2004). Die Modelle sagten vorher, dass einfachere Vorhersagen zu schlechteren Ergebnissen führen, sobald die Umwelt sich verändert, während eine extremere Wahrnehmung von Korrelationen vorteilhaft ist, um die Veränderung zu entdecken.

Übereinstimmend mit meiner Alternativhypothese, dass niedrigere Kapazität zu einfacheren Vorhersagen führt, konnten zwei Experimente zeigen, dass die niedrige Kapazität nur solange von Vorteil ist, bis sich die Umwelt verändert, während sie anschließend nachteilig ist. Übereinstimmend damit konnte ein drittes Experiment zeigen, dass Menschen, die aufgrund der Ablenkung durch eine Zweitaufgabe stärker maximierten, sich nur schwerlich einer veränderten Umwelt anpassen konnten. Der weniger-ist-mehr Effekt hat seinen Preis in einer instabilen Umwelt. Außerdem hat er den Preis, dass Menschen mit geringerer Kurzzeitgedächtniskapazität eher zu falschen Alarmen neigen, was ebenfalls durch das Modell der Hypothese einfacherer Vorhersagen vorhergesagt wurde und in einem vierten Experiment bestätigt wurde.

Folglich ist „probability matching“, oder der Suchprozess, der dazu führt, vermutlich doch nicht so dumm sein, wie es zunächst erscheint. Erstens scheint es eine gute Strategie in den Situationen zu sein, in denen das Individuum nicht allein ist (Gallistel, 1990; Thujisman et al., 1995). Außerdem ist der darunter liegende Suchprozess nach Mustern in vielen Situationen intelligent, weil häufig die Kosten des Verpassens einer nicht-zufälligen Abfolge höher sind als der Preis des Entdeckens eines Musters, wo es eigentlich keines gibt (Lopes, 1982). Es erscheint nur wie ein irrationales Verhalten in Situationen wie der binären Wahlaufgabe, die Bedingungen aufweisen, wie sie außerhalb von psychologischen Labors und Kasinos nur selten vorzufinden sind (Ayton & Fischer, 2004).

Viele Entscheidungen, denen wir gegenüberstehen, sind jedoch weder so einfach wie die binären Wahlaufgaben, noch zeichnen sie sich durch einen solchen Mangel an Informationen aus. Folglich beschäftigte sich Kapitel 2 mit gedächtnisbasierten Entscheidungen in einer komplizierteren Umwelt, in der es mehrere Informationen gab, die für die Entscheidungen eine Rolle spielten. Die Arbeit in diesem Kapitel wurde durch die Arbeit von Bröder und Schiffer (2003b, 2006) inspiriert, die die Idee der Gedächtnissuche bei Entscheidungen in einem Lernparadigma untersuchten. Hier mussten Teilnehmer Informationen über einige Objekte lernen, bevor sie diese Objekte paarweise miteinander bezüglich eines Kriteriums vergleichen sollten. Während der Entscheidungsphase waren keine Informationen über die Objekte sichtbar, so dass die Teilnehmer diese aus dem Gedächtnis abrufen mussten.

Übereinstimmend mit den Befunden aus Kapitel 1, dass eine begrenzte Gedächtniskapazität zur Verwendung einfacherer Strategien führt, zeigen die Resultate von Bröder und Schiffer (2003b, 2006), dass die Notwendigkeit, Informationen aus dem Gedächtnis abzurufen, den Gebrauch von einfachen Entscheidungsstrategien förderte, insbesondere wenn die Gedächtnisbelastung hoch war.

Das Untersuchen von gedächtnisbasierten Entscheidungen weist jedoch das Problem auf, dass der Suchprozess nach Informationen nicht direkt beobachtbar ist – im Gegensatz zu Entscheidungen, bei denen die Information am Bildschirm abgerufen werden kann. Zur Klassifikation von Menschen hinsichtlich der unterschiedlichen Strategien, die sie vermutlich verwendeten, haben Bröder und Schiffer (2003b, 2006) daher nur die Entscheidungen selbst berücksichtigt. Um diese Ergebnisse konvergent zu unterstützen, habe ich Antwortzeiten in den Daten von fünf Experimenten von Bröder und Schiffer reanalysiert. Ein neues Experiment war zusätzlich notwendig, um eine konfundierte Variable zu entwirren.

Die Idee hinter den Antwortzeitanalysen war, dass die unterschiedlichen Strategien unterschiedliche qualitative Vorhersagen über das zu erwartende Muster an Antwortzeiten machen. Eine nichtkompensatorische lexikographische Strategie wie „Take The Best“ (TTB) macht die Vorhersage, dass Menschen die Informationen in der Reihenfolge ihrer Wichtigkeit (Validität) verarbeiten und die Suche beenden, sobald eine Information zwischen den zwei Objekten diskriminiert. Für Menschen, die TTB anwenden, sollten Vergleiche zwischen Objekten, in denen bereits die wichtigste und daher erste Information diskriminiert weniger Zeit in Anspruch nehmen, als Vergleiche, in denen erst die 2., 3.

oder 4. Information diskriminiert. Kompensatorische Strategien wie zum Beispiel „Franklins Rule“ (FR) oder „Dawes Rule“ (DR) gehen dagegen davon aus, dass immer alle Informationen berücksichtigt werden. Daher ist ein Anstieg der Antwortzeiten wie bei TTB für Personen, die diese Strategien verwenden, nicht zu erwarten.

Die vorgefundenen Antwortzeitmuster passten tatsächlich genau zu den auf den Entscheidungen beruhenden Strategieklassifikationen: Benutzer von TTB zeigten die größte Zunahme der Antwortzeiten in Abhängigkeit davon, wie viele Informationen TTB für eine bestimmte Entscheidung benötigte, wohingegen es für Benutzer von DR oder FR keine oder nur eine schwache Zunahme der Antwortzeiten gab. Außerdem waren Benutzer der komplizierteren FR insgesamt langsamer als Benutzer von DR, da FR zusätzlich zu dem Addieren von Informationen auch noch das Gewichten von Informationen erfordert. Am schnellsten waren Personen, die anscheinend geraten haben und somit gar keine Informationen berücksichtigen mussten. Einige Menschen wendeten eine nichtkompensatorische Strategie an, die TTB ähnlich ist, die aber Informationen nicht in der Reihenfolge der Wichtigkeit verarbeitet, sondern in der Reihenfolge, in der dies am einfachsten ist, nämlich in der Reihenfolge, in der die Information gelernt wurde.

Das heißt, diese Menschen nutzten die Flüssigkeit, mit der sie Informationen abrufen konnten, um diese zu ordnen. In diesem Experiment hing die Flüssigkeit des Abrufens von Informationen nur von der Lernreihenfolge ab und war daher nicht informativ. Außerhalb des Labors jedoch ist die Flüssigkeit des Abrufens von Informationen informativ, weil sie stark davon abhängt, wie häufig und wie kürzlich man dieser Information in der Umwelt begegnet ist (J. R. Anderson & Schooler, 1991). In vielen Domänen tauchen Objekte mit größeren Kriteriumswerten häufiger in der Umwelt (z.B. den Medien) auf. Dies gilt beispielsweise für geographische Objekte wie Städte (Goldstein & Gigerenzer, 2002), aber auch für politische Parteien (Marewski, Gaissmaier, Dieckmann, Schooler & Gigerenzer, 2005). Damit übereinstimmend konnten Schooler und Hertwig (2005) zeigen, dass die Flüssigkeit des Erkennens eines Objektes (hier: einer deutschen Stadt) dazu benutzt werden kann, um die Größe dieser Stadt zu erschließen.

In Kapitel 3 erweiterte ich die Idee, dass Flüssigkeit informativ ist, auf die Ebene von Informationen über Objekte. Insbesondere adressierte ich die Frage, ob Menschen die Flüssigkeit, mit der ihnen Informationen über Objekte in den Sinn kommen, prinzipiell verwenden könnten, um diese Informationen zu ordnen. Die Reihenfolge, mit der Informationen berücksichtigt werden ist für Strategien wie TTB von großer Bedeutung, da

ein großer Teil des Erfolges dieser Strategien auf eine gute Reihenfolge zurückzuführen ist. Diese Vorbedingung von TTB macht daraus in den Augen einiger Forscher eine schwierigere Strategie, als es auf den ersten Blick erscheint (Juslin & Persson, 2002).

Um zu erheben, wie genau und wie flüssig Probanden verschiedene Informationen über verschiedene Objekte parat hatten, führte ich ein Experiment durch, in dem ich sie über deutsche Städte befragte. Die Fragen bezogen sich auf insgesamt elf verschiedene Attribute, die jede von 20 ausgewählten Städten entweder aufwiesen oder nicht (z.B., ob eine Stadt einen Flughafen hat oder nicht). Die Annahme war, dass die Probanden diese Fragen umso schneller beantworten konnten, je flüssiger die Informationen vorhanden waren.

Die Ergebnisse dieses Experimentes zeigten, dass die Flüssigkeit, mit der den Probanden Informationen über die Städte in den Sinn kamen, in der Tat informativ ist. Positive Informationen (d.h., das Vorhandensein bestimmter Attribute, wie beispielsweise eines Flughafens) hatten sie schneller parat für größere Städte, vermutlich, weil ihnen Informationen über größere Städte häufiger in ihrer Umwelt begegnen. Für negative Informationen (d.h., für das Fehlen eines Attributs), war dies genau umgekehrt: Negative Informationen konnten flüssiger für kleinere Städte abgerufen werden. Zusätzlich war der Abruf falscher Informationen (d.h., das falsche Beantworten einer Frage) im Durchschnitt langsamer als der Abruf korrekter Informationen, was mit vielen Befunden der Gedächtnisliteratur übereinstimmt (z.B., Ratcliff & Smith, 2004).

Diese Ergebnisse könnten nützlich sein für Strategien, die Informationen nach der Flüssigkeit ordnen. Wenn eine Person gebeten wird, vorherzusagen, welche von zwei Städten größer ist, werden dieser Person im Durchschnitt schneller positive Informationen über die Stadt in den Sinn kommen, die tatsächlich größer ist. Ferner werden ihr schneller negative Informationen über die Stadt in den Sinn kommen, die tatsächlich kleiner ist. Zudem geraten inkorrekte Informationen in der Rangfolge nach hinten, da diese langsamer abgerufen werden, was eine Strategie davor schützen könnte, falsche Informationen überhaupt zu berücksichtigen.

In einem nächsten Schritt habe ich die erhobenen Flüssigkeiten des Informationsabrufs dazu verwendet, um zu simulieren, wie gut einfache Strategien sein können, die Informationen nach dieser Flüssigkeit ordnen. Die Aufgabe bestand darin, in allen möglichen Paarvergleichen zwischen den 20 im Experiment verwendeten Städten vorherzusagen, welche der beiden Städte die größere ist. Das Grundprinzip hinter all

diesen Strategien ist, dass sie Evidenz für oder gegen die Städte akkumulieren, in der Reihenfolge der Flüssigkeit (beginnend mit der flüssigsten Information), bis eine Entscheidungsschwelle erreicht ist (z.B., wenn drei Informationen dafür sprechen, dass eine Stadt die größere ist). Diese Strategien wurden mit Strategien verglichen, die strukturell identisch sind, außer dass sie Informationen in einer zufälligen Reihenfolge verwendeten. Ferner wurden sie auch mit anderen Entscheidungsmodellen wie TTB oder multipler Regression verglichen.

Die Simulationen zeigten, dass das Ordnen der Informationen nach Flüssigkeit sehr erfolgreich ist: Strategien, die dies taten waren wesentlich erfolgreicher als Strategien, die die Informationen in einer zufälligen Reihenfolge verwendeten, insbesondere dann, wenn insgesamt nur wenige Informationen berücksichtigt wurden. Der Vergleich mit den anderen Modellen ergab, dass die Strategien, die die Informationen nach Flüssigkeit ordneten, genauso erfolgreich wie oder sogar erfolgreicher waren als TTB, zum Teil sogar oder unter Verwendung von weniger Informationen! Teilweise konnten diese Strategien sogar mit multipler Regression mithalten – und all das, ohne auch nur die geringste Ahnung von der Wichtigkeit der Informationen zu haben. Menschen können hier folglich einfach die Struktur der Umwelt, wie sie sich in ihrem Gedächtnis widerspiegelt, die Arbeit machen lassen und sich auf das verlassen, was ihnen als erstes in den Sinn kommt.

Im Allgemeinen hoffe ich, dass meine Dissertation dazu beigetragen hat, zu zeigen, dass Menschen weder vollständige Informationen noch unbegrenzte Zeit brauchen, um gute Urteile zu fällen. Die Strategien, die sie verwenden, sind angepasst an die Strukturen der Umwelt und an den menschlichen Verstand und können daher erfolgreich und einfach zugleich sein. Das Gedächtnis kann helfen, indem es die Art und Weise formt, in der Menschen nach Informationen suchen, die sie zuvor gespeichert haben. Es kann die Suche in Richtung der nützlichen Informationen lenken und hindert uns daran, zu viele Informationen zu suchen, die nicht notwendig oder sogar schädlich sein könnten. Einen ähnlichen Gedanken hatte bereits Williams James (1890): Aus seiner Sicht, die ich voll und ganz teile, sind die Beschränkungen des Gedächtnisses ein wichtiger Teil der überlebenswichtigen Selektivität unseres Verstandes.

Dipl.-Psych. Wolfgang Gaissmaier
Holsteinische Str. 50
12163 Berlin

Erklärung

Hiermit versichere ich, dass ich die vorgelegte Arbeit „The Mnemonic Decision Maker: How Search in Memory Shapes Decision Making" selbständig verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht verwendet. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Die Arbeit ist nicht als Ganzes veröffentlicht. Kapitel 1 ist eine erweiterte Version eines Artikels, der im Journal of Experimental Psychology: Learning, Memory and Cognition erschienen ist (Gaissmaier, Schooler, & Rieskamp, 2006). Kapitel 2 ist eine erweiterte Version eines Artikels, der bei Psychonomic Bulletin and Review erschienen ist (Bröder & Gaissmaier, 2007). Ferner sind Teile sämtlicher Kapitel bei verschiedenen Konferenzen vorgestellt worden (siehe Curriculum Vitae). Kapitel 3 (mit Teilen von Kapitel 2) ist hierbei 2006 mit dem Brunswik New Investigator Award ausgezeichnet worden.

In näherer Zukunft ist vorgesehen, bisher unveröffentlichte Teile der Dissertation in überarbeiteter und erweiteter Form zur Veröffentlichung bei Fachzeitschriften einzureichen. Drei Artikel sind in diesem Zusammenhang geplant. Die Experimente 3 und 4 aus Kapitel 1 bilden die Grundlage für einen Artikel, bei dem Lael Schooler mein Koautor sein wird. Der Abschnitt „1.11 Probability Matching Reconsidered" ist Teil eines umfangreicheren Review-Artikels über Probability Matching. Kapitel 3 ist Ausgangspunkt für die theoretische Entwicklung gedächtnisbasierter ökologisch rationaler Strategien. Hierbei werden meine Koautoren Arndt Bröder, Lael Schooler und Julian Marewski sein.

Alle angeführten Koautoren werden bestätigen, dass ich für das Entstehen der Kapitel der Hauptverantwortliche war.

Wolfgang Gaissmaier
Berlin, 5. Januar 2007