CHAPTER 2
Recognition: Inference by Partial and Systematic Ignorance

2.1 On the Psychology of the Recognition Heuristic: Retrieval Primacy as a Key Determinant of its Use

What distinguishes man from beast? This question—perhaps as old as human thought—has prompted many candidate demarcational qualities including man's possession of a soul, knowledge of his own mortality, his malleability, his free will, his language faculty, his abilities of conscious foresight and of genuine, disinterested, true altruism (e.g., Mead, 1935; Dawkins, 1989; Dennett, 1996). At the root of several of these qualities that purportedly distinguish us from all other creatures is man's cognitive capacity for conscious reasoning. As the philosopher Brandom (1994) put it: "Reason is as nothing to the beasts of the field. We are the ones on whom reasons are binding, who are subject to the peculiar force of the better reason" (p. 5).

If the capacity for higher-order cognitive processes is what bestows uniqueness on us, it is not without irony that cognitive psychologists have painstakingly demonstrated the myriad bounds on the very higher-order processes that afford us to excel all other earthly species (e.g., Cowan, 2001; Kahneman, Slovic, & Tversky, 1982; Wason, 1983). The puzzle increases when at the same time it appears as if evolution has hidden—behind the panels of consciousness—a fantastically complex machinery that renders possible elementary processes such as perception, motor coordination, and object tracking, to name just a few. To wit, although evolution appears to have equipped humankind with prodigious processing capacities for seemingly *elementary* processes—the very ones we are likely to share with other animals—it turned strangely stingy when at last it reached the one crowning faculty distinguishing us from them: our ability for higher-order cognitive processes.

But why would evolution have left us with such a "limited-capacity information processor" (Payne, Bettman & Johnson, 1993, p. 9)? Of course, there are several possible reasons, some more obvious than others. One is the considerable cost involved in growing the tissue and in maintaining the metabolism of a large, high-energy expending brain (Martin, 1983). Another is the counterintuitive adaptive benefit of cognitive bounds (e.g., Kareev, 2000; Hertwig & Todd, 2003). Still another, more recently expounded reason (put forth by Gigerenzer, Todd, & the ABC Research Group, 1999) is that even a limited capacity for higher-order cognitive reasoning does not need to stand in the way of enormous intellectual feats. The human mind may have come to use simple cognitive strategies that, in turn, are capable of co-opting automatic and complex evolved (or learned) abilities, to allow the conscious machinery to stay lean.

To illustrate this argument, consider how an experienced baseball player catches a fly ball. Does he do it by subconsciously solving a set of differential equations (as Dawkins, 1989, p. 96, suggested)? Rather than running quickly and in a straight line to the site where the ball is expected to touch the ground—which would be expected if players, consciously or unconsciously, calculated the trajectory—players sometimes trot, or they run toward the ball in arcs, as studies with baseball outfielders show (Shaffer, Krauchunas, Eddy & McBeath, 2004). To wit, players do not seem to calculate the ball's trajectory. But how else do they manage to position themselves to catch the ball? The findings of experimental studies suggest that experienced players rely on several heuristics; one of them is the *gaze heuristic* (e.g., McLeod & Dienes, 1996; Shaffer et al., 2004). The heuristics works in situations in which a ball is already high up in the air, as follows: *Fixate the ball, start running, and adjust your running speed so that the angle of gaze remains constant.*

A player employing this heuristic does not need to measure wind, air resistance, spin, or other variables that determine a ball's trajectory. He can afford to ignore every piece of causal information because the relevant information is encapsulated in one variable: the angle of gaze (i.e., the angle between the eye and the ball, relative to the ground). Although a player who uses this simple and frugal heuristic will not be able to compute the coordinates at which the ball will thud to the ground, the heuristic will nevertheless guide the player to this location (see McLeod & Dienes, 1996, for further details). Importantly, the gaze heuristic can be so effortless because it succeeds in co-opting evolved capacities such as humans' ability to track a moving object against a noisy background. Babies no older than two months can already hold their gaze on moving targets (Rosander & Hofsten, 2002). The ability of object tracking is highly complex and no computer program exists today that manages this ability as well as the human mind does.

Thus one parsimonious answer as to why the human mind can afford to make do with myriad bounds in its cognitive capacities is that it recruits simple strategies that exploit and co-opt complex evolved capacities. As a result, conscious machinery can be kept clear from processes that were they not executed automatically, would impose enormous computational demands. The topic of this chapter is a key example of a simple strategy co-opting a complex capacity. The *recognition heuristic* (Goldstein & Gigerenzer, 2002) hinges on the vast, sensitive, and reliable capacity for recognition. Arguably the most frugal within the program of fast and frugal heuristics (Gigerenzer et al., 1999), the recognition heuristic makes an inference from systematic patterns of existing and missing knowledge. In what follows, I describe the heuristic, the capacity it exploits, and the controversial thesis that it gives rise to noncompensatory inferences.

The Recognition Heuristic: Co-opting an Evolved Capacity

You are a contestant on the ABC show "Who wants to be a millionaire." As your final $1 million question Regis Philbin asks you: Which of the following two musicians has as of today sold more albums in the U.S.: George Strait or Billy Joel? What is your answer? If you are American, then the question may strike you as quite tricky. You may, for instance, remember that throughout his career pop legend Billy Joel has won numerous Grammy Awards, was inducted into the Rock and Roll Hall of Fame, and has released several Top 10 albums. At the same time, you may also think of the many platinum albums that country music legend George Strait has earned, not to mention his many American Music Awards and Academy of Country Music honors. If the choice were tough for an American who happens to know all these facts, how difficult would it be for a European, say, a Swiss, who in all likelihood has never heard of George Strait (93% of students at the University of Basel did not recognize his name; Herzog, 2005), let alone his many achievements?

Yet, could it be that a less knowledgeable Swiss contestant on the show may, paradoxically, be more likely to hit on the right answer than her clued-up American counterpart? More generally, is it possible that people who know less about a subject nevertheless make more correct inferences than their better-informed counterparts? Indeed, it is possible. If the less-informed person—say, the Swiss facing the Billy Joel vs. George Strait question—exploited her systematic ignorance by using the recognition heuristic, she would answer the question correctly: If you recognize the name of one artist but not the other, then infer that the recognized artist has sold more albums. The clued-up American contender cannot use this heuristic, because she has heard of both artists. Ironically, she knows too much to be able to take advantage of the recognition heuristic.

As Goldstein and Gigerenzer (2002) suggested, the recognition heuristic is useful when there is a strong correlation—in either direction—between recognition and the criterion (for simplicity, let us assume henceforth that the correlation is positive). For a two-alternative choice task, such as choosing between Billy Joel and George Strait, the heuristic can be stated as follows:

> *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.

As with the gaze heuristic, the recognition heuristic can be a simple, one-reason decision making strategy (Gigerenzer et al., 1999) because it feeds on the outcome of an evolved capacity. In this case, it is the capacity for recognition, such as face, voice and name recognition. By co-

opting the capacity for recognition, in itself likely to be a complex ability (e.g., Wallis & Bülthoff, 1999), the recognition heuristic can be humble in its demands on cognitive resources.

The capacity for recognition is often assumed to have played a pivotal role in a number of adaptive problems humans have faced throughout their evolution. One such problem is food selection, where in animal and human ancestral environments it would have been too costly to employ a trial-and-error strategy. For instance, when presented with two diets they have not eaten before, Norway rats avoid the diet that they do not recognize from their neighbors' breath. That is, they appear to be operating on the principle that other rats know what they are eating, and this helps them to avoid a potentially deadly consumption of poisons. In addition, according to Galef, McQuoid and Whiskin (1990), this recognition-based decision mechanism works regardless of whether or not the neighbor rat is healthy when its breath is smelled. That is, other potentially vital information beyond recognition cannot override the recognition information. This noncompensatory status of recognition has also been observed in food selection in humans (Hoyer & Brown, 1990).

The Noncompensatory Status of Recognition Information: Mixed Evidence

The capacity for recognition is often assumed to have played a pivotal role in a number of adaptive problems, ranging from avoidance of strangers (Scarr & Salapatek, 1970) to avoidance of poisonous food. In these evolutionarily important domains recognition is often observed to be used in a noncompensatory way (e.g., Galef, McQuoid, & Whiskin, 1990). In light of its evolutionary history, Goldstein and Gigerenzer (2002, p. 77) referred to recognition as a "primordial psychological mechanism," and proposed that the capacity for recognition is being co-opted for drawing probabilistic inferences in the here and now. The recognition heuristic embodies one mind tool through which this co-optation occurs. Moreover, the same authors assumed that the typically noncompensatory status of recognition information observed in evolutionarily important domains generalizes to probabilistic inferences: "The recognition heuristic is a noncompensatory strategy: If one object is recognized and the other is not, then the inference is determined" (Goldstein & Gigerenzer, 2002, p. 82).

The term *noncompensatory* means that for a decision task that is solved based on probabilistic information—cues or attributes—a choice for an object based on one attribute "cannot be reversed by other attributes of the object," that is, the attributes are not integrated into a single judgment (Elrod, Johnson, & White, 2003, p. 2; see also Payne et al., 1993, p. 29).

Relatedly, the recognition heuristic is noncompensatory in that it does not allow room for the integration of recognition knowledge with other probabilistic cues: It "relies only on subjective recognition and not on objective cues" (Goldstein & Gigerenzer, 2002, p. 82). This does not mean, however, that no other knowledge—such as direct knowledge of the object's criterion value—can override the verdict of the recognition heuristic. This very point and, more generally, the meaning of the term noncompensatory seems to have led to some confusion. I return to this shortly.

Since Goldstein and Gigerenzer (2002) proposed the recognition heuristic numerous studies have demonstrated that recognition or lack thereof is an important piece of information across various inferential tasks.[11] At the same time the assumption that it is used in a noncompensatory way has been vigorously challenged (e.g., Bröder & Eichler, 2006; Newell & Fernandez, in press; Newell & Shanks, 2004; Oppenheimer, 2003; Pohl, in press; Richter & Späth, 2006). Goldstein and Gigerenzer (2002) originally tested this assumption by pitting recognition information against other, conflicting probabilistic cues. Specifically, American students were tested on their ability to infer which was the larger of two German cities. Goldstein and Gigerenzer found that despite the presence of conflicting useful cue knowledge that participants had learned during the experiment (e.g., that a particular recognized city has no soccer team), on average 92% of inferences were consistent with the recognition heuristic, suggesting that recognition knowledge overrode knowledge of objective probabilistic cues (but see Newell & Fernandez, in press, Experiment 1; Richter & Späth, 2006, Experiment 3).

In an inventive set of studies, Newell and Shanks (2004) extended the test of the recognition heuristic to a situation in which participants learned to "recognize" fictional company names (consisting of nonwords—that is, none of the names was recognized before the experiment), which were presented repeatedly to them (Bröder & Eichler, 2006, used a similar methodology). Moreover, the validity of the induced recognition was manipulated. In a subsequent judgment task the participants were to infer which of two companies—one recognized, one unrecognized—had the more profitable stock. To aid their decision, people could purchase additional cues in the form of experts' advice. The validity of the cues (i.e.,

[11] Recognition information has been shown to be used efficiently across a range of inferential tasks such as the prediction of outcomes at sports tournaments (Chapter 2.2; Serwe & Frings, in press; Snook & Cullen, 2006), political elections (Marewski, Gaissmaier, Dieckmann, Schooler, & Gigerenzer, 2005) and the estimation of demographic, geographic, and biological quantities (Pohl, in press; Reimer & Katsikopoulos, 2004; Richter & Späth, 2006).

recognition and the recommendations of three advisors) was learned through feedback over the course of the experiment. Consistent with the recognition heuristic, in the majority of choices the recognized company was chosen to be more profitable (88%; see Newell and Shanks' Table 2). In addition, recognition was frequently (68% of all cases) the only cue used (i.e., no further information was purchased). However, this was only so when recognition was the most valid cue. When it was the cue with the lowest validity, most participants (64%) purchased additional information and, based on the experts' advice, a substantial proportion picked the stock they did not recognize (in 38% of cases). Newell and Shanks concluded: "We found little evidence suggesting that recognition is treated any differently from other cues in the environment" (p. 932). In their view, recognition is usually integrated with other available cue knowledge (see also Richter & Späth, 2006). Based on the observation that knowledge of additional probabilistic cues—participants learned them during the experiment—affected the use of (induced) recognition, Bröder and Eichler (2006) arrived at the same conclusion.

Thus, Newell and Shanks' (2004) findings appear to suggest that people stray from the use of recognition as described in the model of the recognition heuristic. How representative, however, is the context in which their participants found themselves in these studies? They knew that recognition was inferior to all other accessible cues. They knew the context in which they learned to recognize an object. This knowledge suggested no association between recognition and the criterion. Outside the laboratory one is rarely so clairvoyant. For instance, one is typically not able to pin down and discern between the various contexts in which one may have previously encountered the names of cities, Goldstein and Gigerenzer's (2002) domain of inference. Thus, Newell and Shanks' (and Bröder & Eichler's, 2006) results may in fact demonstrate that an induced sense of recognition—which can be unmistakably traced to one source, the experiment—may not give rise to the same use of recognition as would a naturally evolved sense of recognition. The latter typically cannot be traced exclusively to one specific source. Such an interpretation of their results also conforms with evidence indicating that in making inferences people appear to rely less on subjective assessments of memory (e.g., processing fluency) when they can attribute this memory to the experiment than when such an explicit attribution is impossible (e.g., Jacoby, Kelley, Brown, & Jasechko, 1989; Oppenheimer, 2004; Schwarz et al., 1991).

The relation of recognition and other knowledge was also the subject of Oppenheimer's (2003) investigation. Unlike Newell and Shanks' (2004) studies, his involved recognition that partly evolved outside the laboratory. Specifically, he presented Stanford students with pairs of well-known and fictitious cities. Their task was to choose the larger one. The well-known cities were carefully selected such that participants either knew that the city they recognized was relatively small (e.g., Sausalito; Experiment 1), or they knew that their ability to recognize a city was due to factors other than its size (e.g., Chernobyl; Experiment 2). In both contexts, Oppenheimer found that recognition information was overruled. The unrecognized fictitious cities were systematically inferred to be *larger* than the recognized cities (i.e., > 50% of the time).

Suspending the recognition heuristic when one explicitly knows that a city is very small, however, does not conflict with the model of the heuristic. In answering questions such as which of two cities is larger, it is plausible to assume that the mind attempts a direct solution by retrieving definitive knowledge about the criterion that gives rise to a *local mental model* (LMM; Gigerenzer, Hoffrage, & Kleinbölting, 1991). In general, an LMM can be successfully constructed if (a) precise figures can be retrieved from memory for both alternatives (e.g., cities), (b) nonoverlapping intervals of possible criterion values can be retrieved, or (c) elementary logical operations can compensate for missing knowledge (e.g., if one city is the largest or the smallest in the set, then any other will by definition be smaller or larger, respectively). An LMM represents a local and direct solution. No use of the probabilistic cue–environment structure is made and merely the presented alternatives and their criterion values are taken into account.[12] According to Gigerenzer et al., only if no LMM can be constructed, will inductive inferences involving probabilistic cues need to compensate for missing direct knowledge. The recognition heuristic is meant to be one model for such an inductive inference, where "the criterion is not immediately accessible to the organism" (Goldstein & Gigerenzer, 2002, p. 78; see also Gigerenzer & Goldstein, 1996). Returning to Oppenheimer's (2003) results, one interpretation is that his students did not employ the recognition heuristic because they succeeded in constructing an LMM, for instance, by assuming that Sausalito is so small that one can safely deduce that the other city, even if not recognized, is larger. Pohl's (in press) results can be interpreted similarly.

[12] Because people's knowledge is imperfect, it is not guaranteed that LMMs yield accurate solutions. Moreover factors such as forgetting and fluctuations in retrieval performance can result in intervals of criterion values rather than precise point estimates.

Across four studies, he found the choice of a recognized object depends, sometimes to a great extent, on whether this choice proves to be correct or incorrect. This contingency would arise if direct (valid) criterion knowledge were available. Finally, Richter and Späth's (2006; Study 1) findings are also consistent with the view that criterion knowledge mediates the use of the recognition heuristic.[13]

Newell and Shanks' (2004) and Oppenheimer's (2003) studies can thus be seen as identifying two important situations in which people clearly do not use the recognition heuristic: First, the heuristic appears not to be triggered or is overruled when recognition knowledge did not evolve naturally and/or when recognition can be traced to one source that is dissociated from the criterion variable. Second, the heuristic, as an inductive device, will only be used if a direct solution fails. Under these conditions, the evidence suggests that people's inferences are not determined by recognition but by information beyond recognition. But these boundary conditions do no warrant the conclusion that recognition information is treated on a par with any other probabilistic information. I submit the thesis that recognition information—independent of its precise confluence with direct and probabilistic knowledge—is not just like "any other" probabilistic cue (Newell & Shanks, 2004, p. 928). Due to its mnemonic properties, recognition has an exceptional status. To appreciate this thesis, let us turn next to research on recognition memory.

Recognition Information: First on the Mental Stage

For more than 30 years, memory researchers have attempted to elucidate the processes underlying recognition (see Yonelinas, 2002, for an overview). Although there is ongoing debate as to whether recognition judgments are based on a single, global-matching process (see Clark & Gronlund, 1996, for a review) or can better be described in terms of a dual-process account (e.g., Jacoby, 1991), there is consensus that two different kinds of information contribute to recognition.[14] One is a global sense of "familiarity", "which is associated with fluent conceptual and perceptual processing, stimulus similarity, and a vague, source-nonspecific feeling of remembrance." The other is the recollection of associative information, including further

[13] In the decision task of Richter and Späth's Experiment 1, participants judged which of two animal species has a larger population. Additional knowledge was assessed by asking participants to indicate whether a species is an endangered one. As endangered species have by definition a small population size, this knowledge represents criterion knowledge.

[14] See Gronlund and Ratcliff (1989) and Clark and Gronlund (1996) for possible accounts of the contribution of these two kinds of information in (modified) global matching models.

knowledge associated with the object and the "conscious retrieval of veridical episodic information from an earlier encounter with a stimulus and gives rise to a feeling of reliving a past event" (Higham & Vokey, 2004, p. 714). For instance, in order to discriminate between a word and a dissimilar nonword (as in a lexical decision task), one can rely exclusively on a global sense of familiarity. This global information, however, will not suffice if a person ought to discriminate the phrase "Bill hit John" from "John hit Bill" when "John hit Bill" was originally studied. Here associative knowledge (e.g., episodic knowledge) has to be brought to bear on the task (Gronlund & Ratcliff, 1989).

A key difference between familiarity and associative information is that familiarity enters the mental stage earlier than does associative information (Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994; McElree, Dolan, & Jacoby, 1999; Ratcliff & McKoon, 1989). This retrieval advantage is generally interpreted to indicate that familiarity represents an automatic form of memory, whereas associative information is rendered available through an intentional, slow and effortful recollection process (Atkinson & Juola, 1974; Jacoby, 1991; Mandler, 1980).[15] Familiarity-based recognition judgments are very fast indeed. For instance, in studies of lexical decisions, when global familiarity allows for a correct response, near perfect performance is reached at around 500 ms after stimulus presentation (e.g., Wagenmakers, Zeelenberg, Steyvers, Shiffrin, & Raaijmakers, 2004; Wagenmakers, Steyvers, Raaijmakers, Shiffrin, von Rijn, & Zeelenberg, 2004).

To summarize, there is evidence that recognition based on a global sense of familiarity is generated automatically (thus requiring little to no cognitive effort), cannot be suppressed (that is, one cannot intentionally stop the recognition process midstream), and precedes other information. These properties distinguish familiarity-based recognition from other pieces of knowledge. Unlike recognition information, additional knowledge—for instance, in terms of probabilistic cues such as whether a given German city has a soccer team or Billy Joel has won numerous Grammy Awards—requires explicit, effortful retrieval and thus is not as rapidly available as recognition. In this sense, recognition is first on the mental stage, ready to enter inferential processes when other the pieces of knowledge (cues) still await retrieval. Henceforth, I refer to these unique properties as the *retrieval primacy* of recognition information.

[15] Furthermore, recollection-based, but not familiarity-based, recognition judgments are susceptible to encoding, division of attention, effects of aging. Context information decays more quickly than familiarity (Skurnik, Yoon, Park, & Schwarz, 2005).

Predictions

The notion of recognition primacy has testable implications. In what follows, I elaborate these implications in terms of three predictions.

> *Prediction 1:* Shorter response times for recognition-based inferences. Inferences that agree with the recognition heuristic require less response time than choices that are inconsistent with the recognition heuristic.

This prediction is derived in the following way: Goldstein and Gigerenzer (2002) described the recognition heuristic as being based on the simple discrimination between "novel and […] previously experienced" objects (p. 77). To render this discrimination possible, typically, global familiarity information suffices and no episodic knowledge is required. As global familiarity information is available almost immediately, inferences based on familiarity-driven recognition will be made expeditiously. In contrast, inferences inconsistent with the recognition heuristic need to rely on information beyond recognition (unless they are produced by mere guessing), such as associated information (e.g., source information) or probabilistic cues. The latter typically require effort and time for retrieval. Hence, such inferences will, on average, require more time than inferences consistent with the recognition heuristic.

Prediction 1 has an interesting corollary: The longer it takes to arrive at a response, the more likely the response will not agree with the recognition heuristic. In other words, with increasing response times there will be a monotonic drop in the proportion of inferences consistent with the recognition heuristic. This regularity follows from the fact that the more time elapses, the more knowledge beyond recognition (if available) can be retrieved. Consequently, the longer the response time, the weaker the impact of recognition on the final judgment.

> *Prediction 2:* Time pressure fosters recognition-based inferences. Limited time to make inferences will lead to greater use of the recognition heuristic, and consequently to more inferences consistent with the heuristic.

Prediction 2 is derived as follows: Recognition is assumed to precede the retrieval of other pieces of knowledge such as probabilistic cues. Because recognition is available when other knowledge could not yet be accessed, it will have more impact on the inferences when this process is subject to time pressure.

In order to derive the final prediction, we first need to turn to another key property of the recognition heuristic. The recognition heuristic is domain-specific, that is, its use will only be successful if recognition is correlated with the criterion. The heuristic's attainable accuracy (i.e.,

the percentage of correct inferences) in an environment is indexed by the *recognition validity* α, which can be calculated as follows:

$\alpha = R / (R + W)$,

where R equals the number of correct inferences, and W equals the number of incorrect inferences (across all inferences in which one object is recognized and the other is not).

Is the recognition heuristic used, and if so, how, in environments in which recognition is but a poor predictor for the criterion? According to Goldstein and Gigerenzer (2002), a somewhat high level of recognition validity is not a prerequisite for the use of the recognition heuristic. As long as α is larger than .5, recognition it can be used. In fact, given its purported noncompensatory nature it will be used if one of two objects is recognized and the other is not—even if conflicting and markedly more valid cues could be retrieved. In contrast, Newell and Shanks (2004) suggested that recognition is used just as any other cue. On their account, the issue of how it is used in environments with low α awaits an elegant solution. If the validity of other cues exceeds α, then recognition will either be weighted less (in case of compensatory processing), or less frequently/not used (in case of noncompensatory processing).

The issue of environments with low recognition validity also pertains to the proposed notion of a retrieval primacy of recognition. Given that recognition is assumed to precede the retrieval of other evidence, how can one escape from the risk of too much reliance on recognition when the recognition heuristic cannot be successfully applied? I suggest that in environments in which the correlation between recognition and the criterion is nil to modest, recognition is only selectively exploited. But, given the hypothesized retrieval primacy of recognition, how can a constrained use be implemented? Specifically, I propose three mechanisms (Predictions 3a–3c) that may achieve a constrained use of the recognition heuristic in environments with low α.

Prediction 3a: Threshold mechanism. According to this mechanism, the user of the recognition heuristic is predicted to rely invariably on recognition as long as α exceeds a threshold. If α falls below this threshold, the user will stop employing the heuristic. Such a threshold mechanism is consistent with the observation that mean adherence rates to the recognition heuristic are consistently high (i.e., around 90%) in spite of extremely varied αs in previous studies (see e.g., Reimer & Katsikopoulos, 2004; Serwe & Frings, 2004; Chapter 2.2). It give rise to three testable regularities: First, although our limited knowledge makes it hard to precisely pin down the numerical value of such a threshold, it should be located between .5 and .7, the lowest α investigated as of today (Chapter 2.2). Second, the threshold mechanism implies two clearly distinguishable clusters of adherence rates—one encompassing high adherence rates (users whose α exceeds the threshold) and another one including low adherence rates (users

whose α is below the threshold). Third, there should be a strong positive correlation between individuals' α and their adherence rate.

Prediction 3b: Matching mechanism. According to this mechanism, the user of the recognition heuristic follows it with a probability that matches α, the recognition validity. This mechanism is inspired by the frequent observation of people choosing the more likely of two events with a probability that matches the probability of success of that event. Specifically, when people have to choose between two options *a* and *b*, and *a* leads to a success with probability *p*, and *b* leads to a success with probability $q = 1 - p$, people respond as if they were probability matching. That is, rather than always choosing *a* (i.e., probability maximization), they distribute their responses such that *a* is chosen with a probability of *p* and *b* is chosen with a probability of $1 - p$ (e.g., Gallistel, 1990; Vulkan, 2000).

In the context of the recognition heuristic, such a mechanism means that people match their use of the heuristic to the environment-specific α (see also Harvey & Rawles, 1992). Consequently, the recognized object is chosen to be the larger one with a probability of $p = \alpha$, and the unrecognized object with a probability of $q = 1 - \alpha$. One implication of this mechanism is that for each item pair in which the recognition heuristic is applicable, the proportion of inferences consistent with it should equal α. Like the threshold mechanism, the matching mechanism implies a strong correlation between adherence rates and individuals' α; unlike the former mechanism, it does not imply two clearly distinguishable clusters of adherence rates but a linear relationship between adherence and α.

Prediction 3c: Suspension mechanism. According to this mechanism, the user of the recognition heuristic may suspend the use of the heuristic if knowledge at odds with recognition is available. Such contradictory knowledge can come in different forms including (a) conflicting probabilistic cues with higher validity (Newell & Shanks, 2004), (b) knowledge of invalid source (i.e., if a person realizes that his recognition of an object stems from an invalid source such as the presentation of an object within an experiment) and (c) direct and conflicting knowledge of an object's criterion value (see Oppenheimer, 2003). In this chapter, I focus on the latter. If, for instance, a Stanford university student is asked to judge which city has more inhabitants, Berkeley or Gelsenkirchen (in Germany), the student might zoom in on Gelsenkirchen albeit she does not recognize this German city. The reason is that she knows that Berkeley, with only slightly more than 100,000 residents, is a small city. Therefore, she suspends the recognition heuristic for this specific inference.

Unlike the first two mechanisms, the suspension mechanism implies marked variability in the use of the recognition heuristic across objects and across participants because some objects are more likely to be associated with direct knowledge than others (e.g., students of Stanford

university are likely to know that Berkeley is comparatively small and may put the recognition heuristic aside in all pairs that involve Berkeley), and some people have direct knowledge where others lack it. Consequently, the suspension mechanism may obliterate a strong link between people's *a*s and their recognition heuristic adherence, a link obligatory for the other two mechanisms.

The Environment

I now turn to two studies that tested Predictions 1, 2 and 3. Both studies used variants of the same experimental procedure. Participants were given pairs of infectious diseases, and their task was to choose the more prevalent in each pair (see also Chapter 3.1[16]). Why this domain? First, because it requires the retrieval of knowledge acquired outside the laboratory, thus liberating us from using experimentally induced recognition or artificially created environments. Equally important, Prediction 3 requires the study of an environment in which recognition is of comparatively low validity. Conveniently, such an environment is also appropriate to test Predictions 1 and 2. Both necessitate an environment in which at least some of people's knowledge conflicts with the recognition heuristic, and this is likely in an environment with low recognition validity. The domain of infectious diseases represents such an environment.

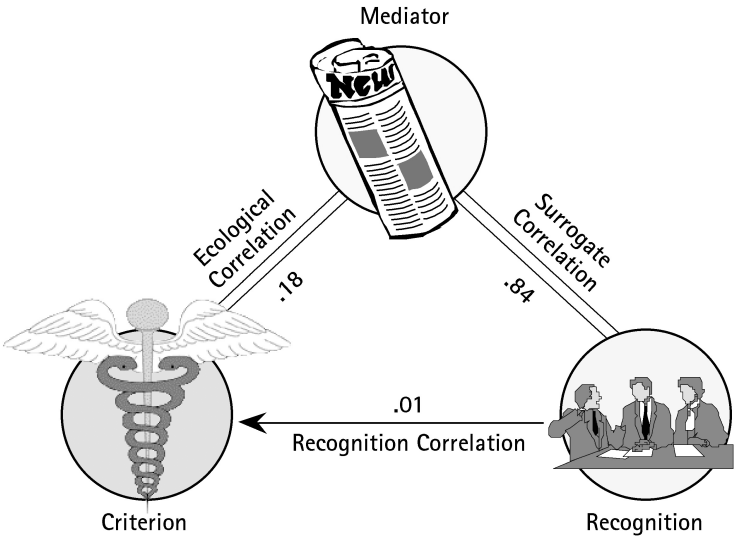

Figure 2.1.1. Ecological analysis of recognition. Recognition is highly correlated with media coverage, whereas both media coverage and recognition are uncorrelated with incidence rates of the infectious diseases. Recognition is a very poor indicator of incidence rates of the diseases, because media coverage, acting as mediator, does not (only) reflect the incidence rates (but possibly also the severity of the disease).

[16] In Chapter 3.1 the recognition heuristic is not investigated directly. However, there was only a modest correlation (r_s = .23) between the incidence rates of the diseases and the frequency with which the infections were mentioned in the media; the latter being a strong predictor of recognition (according to Goldstein & Gigerenzer, 2002).

Figure 2.1.1 depicts the relationships between annual incidence rates of 24 notifiable infectious diseases in Germany, the frequency with which the names of the diseases were mentioned in the media, and *collective recognition* (i.e., the proportion of participants recognizing each infection in Study 1; see Goldstein & Gigerenzer, 2002).[17] The frequencies of mentions in the media, assumed to operate as the mediator between the criterion and recognition, were determined using COSMAS I, an extensive data archive of German daily and weekly newspaper articles.[18] I determined the number of times the names of the 24 infections were mentioned, and rank correlated these numbers with recognition. As Figure 2.1.1 shows, media coverage was highly correlated with recognition (*surrogate correlation*: r_s = .84; p = .001), in line with the assumption that recognition is determined by how often things are mentioned in the media (Goldstein & Gigerenzer, 2002). In contrast, the correlation between the criterion and the mediator, the *ecological correlation*, was weak (r_s = .18; p = .39). Most importantly, the correlation between collective recognition and the infections' incidence rates turned out to be nil (r_s = .01; p = .95). The percentage of participants who recognized the infections did not reflect the actual incidence rate of the diseases. Undeniably, recognition is a poor predictor of the criterion in this environment *hostile* to the recognition heuristic.

Study 1: Does the Recognition Heuristic Give Way to Faster Choices?

If one object is recognized and the other is not, the recognition heuristic can determine the choice without searching and retrieving other information about the recognized object. The reversal of a choice determined by recognition, in contrast, requires retrieval of further information (unless the reversal reflects mere guessing). Hinging on this difference, Prediction 1 states that inferences that agree with the recognition heuristic require less response time than choices that are inconsistent with the recognition heuristic. Study 1 tests this prediction, as well as Prediction 3.

Method

Participants and design. Forty students from the Free University (Berlin) participated in the study (27 women and 13 men, mean age = 24.2 years), which was conducted at the Max

[17] Classified as particularly dangerous, occurrences of these diseases have to be registered. As correct answers, I used statistics prepared by the Federal Statistical Office of Germany and the Robert Koch Institute (e.g., Robert Koch Institute, 2001). To reduce year-to-year fluctuations, the data were averaged across four consecutive years (1997-2000).

[18] COSMAS (Corpus Search, Management and Analysis System) is the largest online archive of German literature (e.g., encyclopedias, books, and newspaper articles; http://corpora.ids-mannheim.de/~cosmas/). The analysis was based on a total of the 1,211 million words.

Planck Institute for Human Development. They were presented with pairs of infectious diseases and asked to choose the infection with the higher annual incidence rate in a typical year in Germany (henceforth *choice task*). They also indicated which of the infections they recognized (henceforth *recognition task*). All people were paid for participating. Half of the participants received a flat fee of €9 (= $11.76 U.S.) and monetary incentive in form of performance-contingent payment. Specifically, they earned 4¢ (= 5¢ U.S.) for each correct choice and lost 4¢ for each wrong one. The other half of participants received a flat fee of €10 (= $13.07 U.S.). Participants were randomly assigned to one of the four conditions of a 2 (recognition test before/after the choice task) × 2 (monetary incentive/no incentive) design, with 10 participants in each condition.

Materials. For the choice task, I used all 24 infectious diseases (see Table 1) and generated all 276 possible pairs, which were presented in 12 blocks (each containing 23 pairs). Both the order in which the 276 pairs of infections appeared and the order of the infections within each pair were determined at random for each participant. The recognition task comprised all 24 infections.

Procedure. After an introductory text explaining the relevance of accurate judgments of the frequency of dangerous infectious diseases, people read the following instructions:

> You are to judge the annual frequency of occurrence of different types of infections in Germany .... Each item consists of two different types of infections. The question you are to answer is: For which of two events is the number of new incidents per year larger?

Participants were presented with the pairs of infections displayed on a computer screen. They were asked to indicate their choice by pressing one of two keys. In addition, they were instructed to keep the index fingers of the right and left hand positioned on the keys representing the right and left element in the pair infections, respectively, for the entire duration of one block, and were encouraged to respond as quickly and accurately as possible (although they were not told that their response times were recorded). The time that elapsed between the presentation of the name of the infections on the screen and participants' keystroke was measured. After conclusion of the choice task, half of the participants took the recognition task. In this task, the 24 infections were presented in alphabetic order on a questionnaire and participants were asked to indicate whether they had heard of the infection before. Half of participants took the recognition test prior to the choice task. On average, the complete session lasted 60 minutes.

Each choice between two infections began with the presentation of a fixation point (a cross in the center of the screen), followed after 1,000 ms by the names of the two infections. The names appeared simultaneously (left and right from the fixation point) and remained on the

screen until a response was given. Participants were informed that once the response key was pressed their choice could not be reversed. After each response, the screen remained blank for 1,000 ms. In order to accustom participants to the procedure, they responded to 10 practice trials.[19]

Results

Before turning to the test of Prediction 1, I describe the obtained inferences in more detail. On average, participants scored 60.9% (*SD* = 5.6) correct. The level of accuracy did not differ significantly across the two incentive conditions ($F[1, 36] = .01$, $p = .94$), and the two orders of the recognition task ($F[1, 36] = 2.34$, $p = .14$). Therefore, I pooled the data for the following analyses. On average, participants recognized 58% (range 35.8%-95.1%) of the 24 infections. Recognition rates are listed in Table 2.1.1.[20] The frequency of recognized infections did not increase significantly when the recognition task succeeded the choice task ($t[38] = 1.31$, $p = .20$). Across all participants and items, the recognition heuristic was applicable in almost half of all pairs (48.5%). Finally, the average recognition validity α was .60 (*SD* = .07). That is, on the individual level there was a modest relationship between disease incidence rate and recognition (chance level = .5). The average knowledge validity β—expressing the accuracy in cases when both diseases are recognized—was .66 (*SD* = .08).

Did the recognition heuristic predict people's inferences? For each participant I computed the proportion of inferences that were in line with the recognition heuristic among all cases in which it could be applied. The proportions of recognition heuristic adherence ranged between 35% and 95%. The mean proportion of inferences in accordance with the recognition heuristic was 62.1% (median 62.7%). Neither task order ($F[1, 36] = .001$, $p = .98$, $\eta^2 = .001$) nor monetary incentive ($F[1, 36] = .66$, $p = .42$; $\eta^2 = .02$) had an effect on proportions of recognition heuristic adherence. The present adherence rate is markedly lower than in Goldstein and Gigerenzer (2002), who found proportions of 90% and higher (in a task involving choosing the larger of two German or U.S. cities). Thus, the investigated environment was indeed one in which people in a substantial portion of their judgments did not obey the recognition heuristic, thus creating a test bed for Prediction 1.

Were inferences in accordance with the recognition heuristic made faster (Prediction 1)? I analyzed the response times by taking choices rather than participants as the unit of analysis. Figure 2.1.2 shows the average response times for inferences consistent and inconsistent with the recognition heuristic at the 25th, 50th, and 75th percentiles of the response-time distribution.

[19] Consisting of 10 randomly drawn pairs of infections, which were replaced for the main task. The responses of the practice trials were excluded from the analysis.
[20] Tularemia was excluded from the following analyses because only one person recognized it.

Consistent with Prediction 1, I found that response times for inference that agreed with the heuristic were consistently shorter at each of the three percentiles than choices conflicting with the heuristic. This result was confirmed by the second analysis, in which the response times were natural log-transformed to reduce the skewness of the data. Figure 2.1.3 compares the average response times for inferences consistent and inconsistent with the recognition heuristic. Again consistent with Prediction 1, inferences that conflicted with the recognition heuristic took longer (*Mdn* = 2,022.5 ms; transformed *M* = 7.7, *SD* = 0.6) than those consistent with the recognition heuristic (*Mdn* = 1,668 ms; transformed *M* = 7.5, *SD* = 0.6; $t(5353) = 10.8$, $p = .001$; Cohen's $d = 0.30$). The latter also took less time than inferences in which the recognition heuristic was not applicable (with *Mdn*s of 2,032 ms and 1,953.5 ms when both diseases were unrecognized and recognized, respectively). Finally, as Figure 2.1.3 also shows, the response times for incorrect inferences were markedly longer than for correct inferences, irrespective of whether or not inferences agreed with the recognition heuristic. This pattern is a typical finding in the memory literature, especially in tasks in which the overall accuracy is low (e.g., Ratcliff & Smith, 2004).

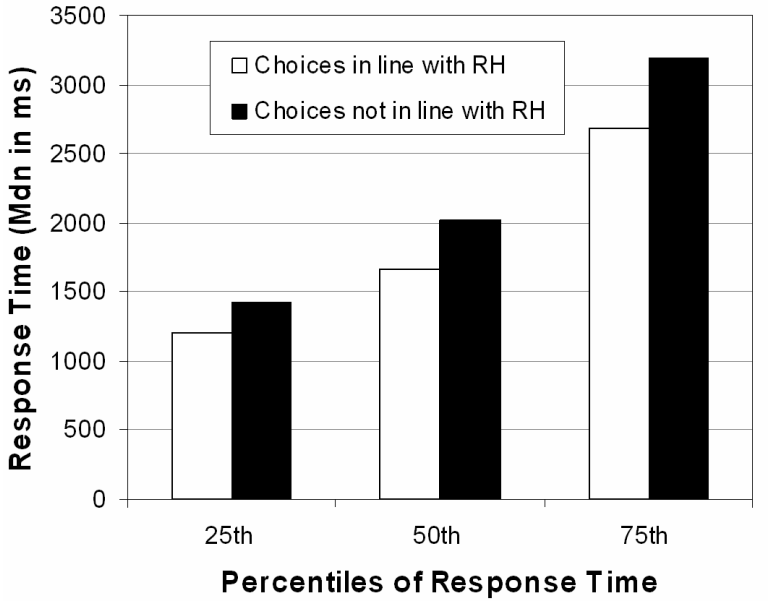

Figure 2.1.2. Distribution of the response times of choices where the recognition heuristic was applicable. The 25th, 50th and 75th percentiles of response times are shown as a function of whether the recognition heuristic was applied or not.

To conclude, in support of Prediction 1, inferences that agreed with the recognition heuristic were made faster than those that went against it. This observation also supports the notion that recognition information outruns other inferential information. The decision not to use the recognition heuristic appears to exact the cost of longer response times.

Table 2.1.1. The 24 infectious diseases used as target events in Studies 1 and 2. Shown are the diseases' annual incidence rate, the proportion of participants recognizing them, adherence to the recognition heuristic (when they were recognized and paired with an unrecognized disease), and the median incidence estimate and proportion for participants with direct knowledge of the prevalence rate obtained in Study 2.

		Study 1 (N = 40)			Study 2 (N = 60)				
	Annual incidence rate	Recognized by % of participants	*M* proportion of choices in line with RH	*n*	Recognized by % of participants	*M* proportion of choices in line with RH	*n*	*Mdn* estimated incidence	% of participants with direct knowledge
Poliomyelitis	0.25	100	0.57	40	100	0.70	59	50	30.0
Diphtheria	1	97.5	0.66	39	98.3	0.70	58	500	18.3
Trachoma	1.75	7.5	0.76	3	13.3	0.49	7	50	5.0
Tularemia	2	2.5	1	1	3.3	0.57	1	50	1.7
Cholera	3	100	0.30	40	100	0.47	59	5	31.7
Leprosy	5	100	0.15	40	100	0.37	59	5	30.0
Tetanus	9	100	0.66	40	100	0.69	59	500	23.3
Hemorrhagic fever	10	20.0	0.76	8	33.3	0.82	19	500	6.7
Botulism	15	22.5	0.63	8	18.3	0.70	10	50	8.3
Trichinosis	22	20.0	0.60	9	23.3	0.67	13	50	5.0
Brucellosis	23	12.5	0.66	5	15.0	0.83	8	50	5.0
Leptospirosis	39	7.5	0.42	3	25.0	0.68	14	50	5.0
Gas gangrene	98	27.5	0.38	11	28.3	0.65	16	50	11.7
Ornithosis	119	7.5	0.54	3	10.0	0.79	5	50	5
Typhoid and paratyphoid	152	87.5	0.46	35	90.0	0.77	53	50	16.7
Q fever	179	12.5	0.37	5	16.7	0.56	9	50	5.0
Malaria	936	100	0.63	40	100	0.59	59	500	26.7
Syphilis	1,514	95.0	0.59	38	100	0.76	59	500	21.7
Shigellosis	1,627	5.0	0.90	2	20.0	0.64	11	50	5.0
Gonorrhea	2,926	95.0	0.74	38	96.7	0.72	57	5,000	18.3
Meningitis and encephalitis	4,019	97.5	0.79	39	91.7	0.88	54	5,000	20.0
Tuberculosis	12,619	100	0.67	40	98.3	0.69	58	500	26.7
Viral hepatitis	14,889	90.0	0.91	36	86.7	0.84	51	5,000	18.3
Gastroenteritis	203,864	85.0	0.97	34	96.7	0.92	57	200,000	26.7

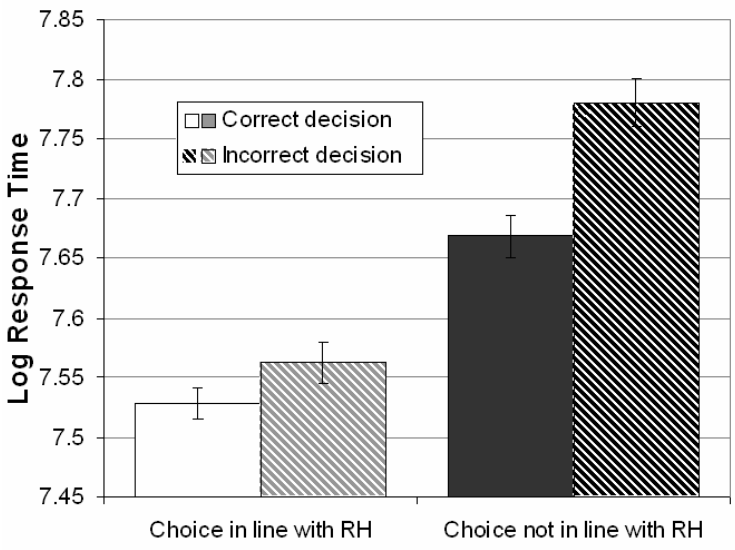

Figure 2.1.3. Response times of choices where the recognition heuristic was applicable as a function of whether the recognition heuristic was applied or not, and of the accuracy of the choice.

Which mechanism explained the restricted use of the recognition heuristic best (Predictions 3a-3c)? As observed earlier, the recognition heuristic accordance is markedly lower in the infectious diseases environment than in other environments previously studied. Therefore, it is possible to investigate which of the proposed mechanisms—the threshold, the matching, and the suspension mechanisms, respectively—is instrumental for the constrained use of the heuristic in this environment. I begin with the threshold mechanism. Here, the average recognition heuristic accordance represents the combination of two clusters of adherence rates: First, the high rates of those who invariably rely on the heuristic because their α (their recognition validity) exceeds the critical threshold; second, the low rates of those who never employ the heuristic because their α is below threshold. Figure 2.1.4 shows the adherence rates for the 40 individual participants. For each participant one bar is shown, representing the proportion of inferences that agreed with the recognition heuristic among all cases in which it could be applied. This distribution of rates does not resemble that implied by the threshold mechanism. Rather than showing two clusters of adherence rates—a cluster of high rates and one of low rates—the actual rates varied continuously between 35.8% and 95.1%.

Figure 2.1.4 also renders possible a test of the matching mechanism. According to this mechanism, the user of the recognition heuristic employs it with a probability corresponding to his or her recognition validity. On an aggregate level, the proportion of choices following the recognition heuristic indeed closely matched the average α: .62 versus .60. In an analysis of individual adherence rates and αs, however, this match proves spurious. The horizontal line in each bar of Figure 2.1.4 represents the person α. Ostensibly, a person's α is not indicative of how often she follows the heuristic. In fact, the correlation between participants' α and

their adherence rate is negligible (r = -.20, p = .23)—a result that disagrees both with the threshold and the matching heuristic.

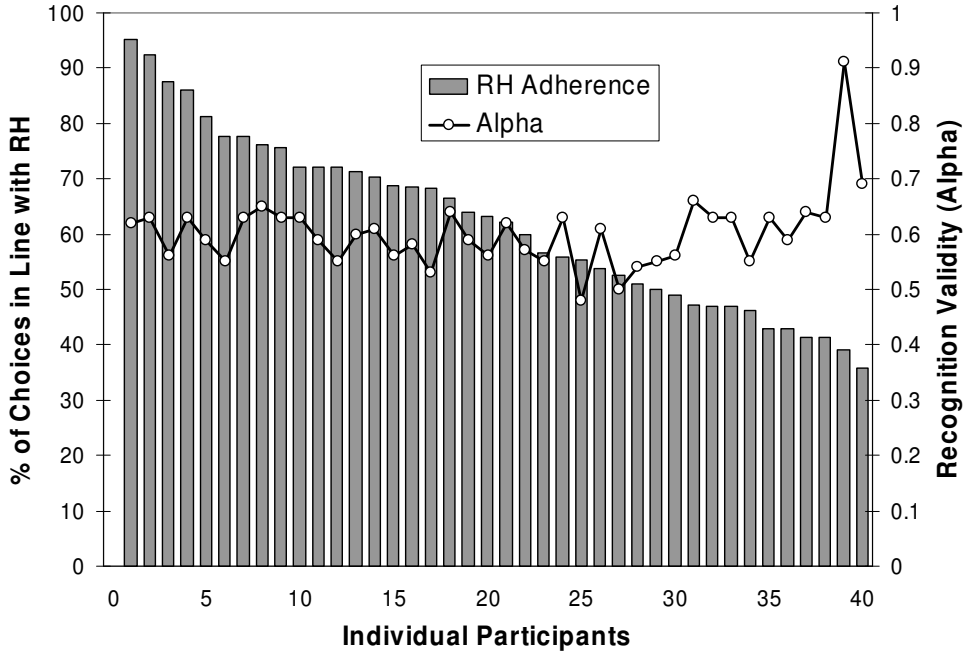

Figure 2.1.4. Adherence to the recognition heuristic and recognition validity by individual participant.

Finally, according to the suspension mechanism, knowledge in conflict with the recognition heuristic can prompt the user to temporarily suspend it. Assuming that objects differ in the degree to which they are associated with such knowledge, the mechanism implies varied adherence rates across objects. To investigate this possibility, I calculated for each infection (averaged across participants) the proportion of cases in which the infection was inferred to be the more frequent one given it was recognized and paired with an unrecognized infection. Figure 2.1.5 plots these proportions, separately for each infection (averaged across participants). Indeed, there were large differences between the infections' adherence rates. Some such as gastroenteritis (.97) and viral hepatitis (.91) were almost invariably chosen over unrecognized ones (if they were recognized). In contrast, infections such as cholera (.30) and leprosy (.15) were mostly inferred to be the less frequent ones. As Table 2.1.1 shows, adherence rates are by no means closely lined up with recognition rates (r = -0.1, p = .98). In other words, commonly recognized infections such as cholera, leprosy, and diphtheria are not necessarily those that yield high adherence rates. What drives people's decisions to distrust recognition? I suspect it is the direct and conclusive criterion knowledge that infections such as cholera and leprosy are virtually extinct in Germany, a possibility that will be explored further in Study 2.

To summarize, I investigated three candidate mechanisms underlying the constrained use of the recognition heuristic in an environment in which the heuristic does not promise to be highly successful. Two of the three mechanisms—the threshold and the matching mechanism—received little support: People did not invariably draw on or suspend the use of the heuristic as a function of whether or not their αs surpassed a threshold (threshold mechanism). Similarly, users of the recognition heuristic did not employ it with a probability corresponding to their αs. Instead, I observed (a) a small correlation between individuals' αs and their heuristic adherence rate, and (b) enormous adherence variability across infections. The latter finding suggests that item-specific conflicting knowledge prompts users not to use the heuristic. Did this knowledge help them to boost their inferential accuracy?

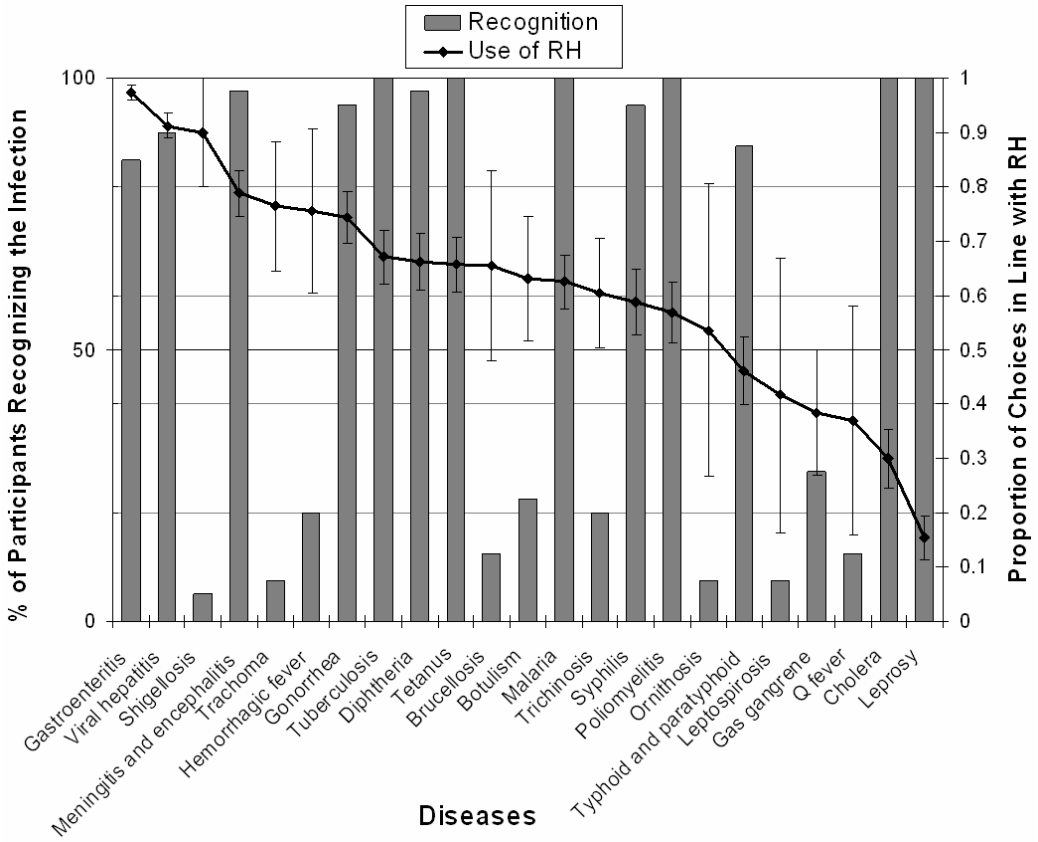

Figure 2.1.5. Object-specific recognition and adherence to the recognition heuristic. For each disease, the average (across participants) proportion of choices in line with the predictions of the recognition heuristic (when the event was recognized and paired with an unrecognized disease) is shown. Tularemia, recognized by only one person, is not shown. Error bars indicate standard errors. The columns represent the percentage of participants who recognized the disease.

Could participants boost accuracy by temporarily suspending the heuristic? Suspending the recognition heuristic temporarily will increase accuracy if it is done in cases in which the heuristic would have arrived at the wrong choice. People's level of accuracy among all cases

in which the recognition heuristic can be applied will then exceed α (i.e., the percentage correct that a user would have achieved if she had invariably used the recognition heuristic whenever applicable). This boost in accuracy, however, will occur only if people have some ability to tell apart cases in which the heuristic yields correct versus false judgments, *and* if the accuracy in the cases in which the heuristic is suspended (i.e., the unrecognized object is chosen) exceeds α.

I tested this possibility as follows: I first turned to the question of whether people can discriminate between cases in which the heuristic arrives at correct versus incorrect inferences. To this end, signal detection theory was used (*SDT*; Green & Swets, 1966). This theory describes a decision maker who must choose between two (or more) alternatives on the basis of ambiguous evidence. This uncertain evidence is summarized by a random variable that has a different distribution under each of the alternatives, here correct versus incorrect inferences when the recognition heuristic is used. The evidence distributions typically overlap, thus sometimes evidence is consistent with both alternatives. The respondent establishes a decision criterion *C* that divides the continuous strength of evidence axis into regions associated with each alternative. If the evidence value associated with an event in question exceeds C, the respondent will conclude: "Following the recognition heuristic leads to a correct inference." Otherwise she will conclude: "Following the recognition heuristic leads to an incorrect inference." Her conclusions result in four types of outcomes: *hits* (use of the recognition heuristic yields a correct inference), *correct rejections* (suspending it yields a correct inference), *misses* (suspending it yields an incorrect inference), and *false alarms* (use of the recognition heuristic yields an incorrect inference).

One measure of a person's ability to tell apart cases in which the recognition heuristic ought and ought not to be used is the distance between the means of the distributions under the two alternatives. If this sensitivity index *d´* is small (i.e., the two distributions overlap considerably), a person will likely fail in boosting her accuracy by temporarily suspending the recognition heuristic. Across all participants, the observed mean *d´* differed significantly from zero ($M = .55$; $SD = .42$; $t[39] = 8.35$, $p = .001$).[21] Participants were thus able to distinguish—although not perfectly—between cases in which recognition would have been an invalid piece of information and those in which it would prove valid. But how good were participants in an absolute sense? In order to further evaluate people's ability to selectively suspend the recognition heuristic when it would result in an error, I devised an index *c*, ranging from -1 to 1. This index separates a person's tendency to use the recognition heuristic (comparable to the bias measure in SDT) from the ability to identify cases in which recognition leads astray, and merely reflects the ability (see Appendix A for its derivation). Index c expresses the actual achieved accuracy for all cases in which the recognition heuristic would have been applicable as a proportion of the maximum accuracy attainable if a person were able to identify all cases

[21] The sensitivity measure *d´* was highly correlated with the non-parametric sensitivity measure *A´* ($M = .66$, $SD = .11$): $r = .98$.

in which recognition leads astray. A value of 0 indicates that the decision to temporarily suspend the recognition heuristic was made purely randomly, thus not reflecting any true ability to discriminate when recognition leads to a correct inference and when it does not. In contrast, a value of 1 means that any single decision to discard recognition (and thus choose the unrecognized object) resulted in a correct inference—controlling for the errors due to a person's tendency to over- or underuse the recognition heuristic. On average, the observed *c* amounted to .30 (*SD* = .20; range from -0.06 to .98) and participants realized 30% of the maximal possible improvement by temporarily suspending the recognition heuristic. The *c* values of 36 participants (90%) were positive, only four participants had negative values.

The *d´* and the c indices demonstrate people's ability—although limited—to discriminate between cases in which recognition was valid and those in which it was invalid. But did this ability actually translate into a higher accuracy? For each person, I calculated the actual accuracy among all items in which the heuristic was applicable. Then, I compared this value to the person's α (the level of accuracy if the person invariably applied the heuristic). Compared to their αs, 24 of 40 participants (60%) managed to boost their accuracy by suspending the recognition heuristic every now and then. The accuracy of 16 participants worsened. On average there was no increase in accuracy: Across all participants, the recognition heuristic would have scored 60.3% (*SD* = 6.7) correct. In comparison, the empirical percentage correct was 60.9% (*SD* = 7.4)—a nonsignificant difference (paired-samples t-test: $t(39) = 0.39$, $p = .70$). In other words, by temporarily suspending the recognition heuristic, people did not succeed in increasing their inferential accuracy beyond the level would have been attainable if they had invariably used the heuristic.

Summary of Study 1

In the first study, I tested Predictions 1 and 3. Consistent with Prediction 1, I observed markedly shorter response times for recognition-based inferences. That is, inferences that were in line with the recognition heuristic proved to require noticeably less response time than those conflicting with it. This finding is consistent with the notion of recognition's retrieval primacy. In contrast with other knowledge, recognition information arrives first on the mental stage and thus has a competitive edge over other pieces of information. Yet, people appear to frequently overrule recognition information in an environment in which there is little to no relationship between recognition and the criterion. Indeed, I found that in such an environment, the use of the recognition heuristic was constrained. Compared to the typical very high adherence rates for the recognition heuristic, I observed an average rate of about 62%. The mechanism that appears to achieve this constrained use is the suspension mechanism (Prediction 3c). Specifically, people appear to decide for individual cases whether or not they will obey the recognition heuristic. Moreover, these decisions are not made arbitrarily but demonstrate some ability to discriminate between cases in which the recognition heuristic would have yielded correct judgments and cases in which the

recognition heuristic would have led astray. This ability, however, does not result in a performance boost because the level of accuracy in cases in which the heuristic was set aside does not exceed α.

Study 2: Does Time Pressure Increase Adherence to the Recognition Heuristic?

Study 1 provided evidence supportive of the notion of recognition's retrieval primacy. Inferences in accordance with the recognition heuristic were made faster than inferences conflicting with the heuristic (Prediction 1). In other words, the decision to set aside recognition information requires extra time. Based on this finding, I now turn to Prediction 2: Bounds on the available response time will increase reliance on the recognition heuristic, and consequently result in a higher rate of inferences agreeing with it. Study 2 tests this prediction. In addition, it further investigates the constrained use of the recognition heuristic in a somewhat 'hostile' environment. In Study 1, it was observed that participants temporarily set aside the heuristic. Across infections, such suspension did not occur randomly but was more pronounced for some infections than for others (see Figure 2.1.5). I now explore what kind of knowledge triggers the decision to suspend the recognition heuristic.

As suggested by the results in Study 1, one possibility is the presence of *direct* and *conclusive* knowledge of the incidence rate of a recognized infection that conflicts with recognition information. For instance, a person may remember that cholera has been virtually eliminated (in Germany). This knowledge suffices for the person to conclude that cholera cannot be more frequent and is likely to be less frequent than any other infection, irrespective of whether it is recognized. In general, I suggest that direct knowledge on the criterion variable will overrule recognition information if the following can be retrieved from memory: (a) non-overlapping criterion intervals, and (b) precise figures (ranks) that in combination with elementary logical operations can compensate for missing knowledge (e.g., a particular infection is known to be the rarest infection, thus by extension any other infection is more frequent).

When Goldstein and Gigerenzer (2002) taught their participants useful information that offered an alternative to following the recognition heuristic, they observed that recognition overruled this evidence. Their additional evidence, however, did not consist in direct and conclusive knowledge of the criterion variable but in terms of a probabilistic cue. Perhaps, this helps to explain why they found such a high rate of adherence to the recognition heuristic in light of conflicting evidence. In Study 2, I examine whether direct knowledge of the criterion variable (that participants bring to the laboratory) can override recognition information.

Method

Participants and design. Sixty students from the Free University (Berlin) participated in the study (41 women and 19 men; mean age = 24.6 years), which was conducted at the Max

Planck Institute for Human Development. As in Study 1, they were presented with 276 pairs of infectious diseases and asked to choose the one with the higher annual incidence rate. Furthermore, each participant indicated which infections (now presented on a computer screen) he or she recognized. Half of participants took this recognition test before the choice task and half after. (Order turned out to have no effect on recognition.) They received an initial fee of €9 (= $11.76 U.S.), and earned 4¢ (= 5¢ U.S.) for each correct answer and lost 4¢ for each wrong answer.

Material. Participants responded to the same 276 infection pairs used in Study 1. In addition, they classified each infection in one of the following six frequency categories: < 1-9, 10-99, 100-999, 1,000-9,999, 10,000-99,999, and > 100,000.

Procedure. Participants read the same introductory text as in Study 1 (see previous *Method* section), after which they were presented with pairs of infections displayed on a computer screen. Time pressure in this choice task was realized as follows (Figure 2.1.6). The pairs of infections were presented sequentially on a computer screen in 12 blocks. Each presentation of one pair of infections began with an acoustic signal (10 ms in length), followed by a second signal (900 ms later) that coincided with the presentation of a small fixation cross in the middle of the screen. Again, 900 ms later, the fixation cross disappeared, and a third signal followed, accompanied by a pair of infections (left and right from the location of the fixation cross). The pair remained on the screen for 700 ms before disappearing. The participants' task was to choose the more frequent infection in each pair, and to indicate their response by pressing one of two keys on the keyboard. They were instructed to respond *as quickly and as accurately as possible*, but not later than a fourth *imaginary* signal, 900 ms after the third tone and the onset of the stimulus presentation (Figure 2.1.6). The reason for using an imaginary signal was to avoid interference of the signal indicating the response deadline with the processing of the stimulus pair (this is a procedure used in research on the lexical decision task; see e.g., Wagenmakers et al., 2004). If a response was markedly delayed (i.e., > 1200 ms after the presentation of the stimulus pair), the message "too late" would appear on the screen, accompanied by an aversive tone. A delayed response reduced the person's income by 4¢ (= 5¢ U.S.). In the recognition task, participants saw the names of the 24 infections one at a time (in random order) on the computer screen. They were asked to decide whether they had heard of the infection and to express their positive or negative answer by pressing one of two keys. At the close of the experiment, every participant was asked to classify each infection in one of six frequency categories and to determine whether this judgment was made on the basis of *certain* knowledge of the criterion variable.

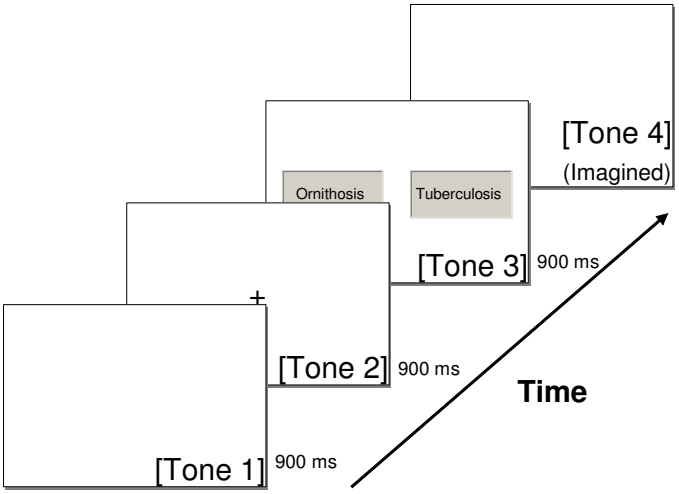

Figure 2.1.6. Induction of time pressure in Study 2.

In order to acquaint participants with the procedure in the choice task, they underwent ten practice trials. Each trial consisted of a pair of arrows (">" and "<", randomly ordered). The task was to indicate within the time limit whether the ">"-arrow was shown on the left or right side of the screen. In a second block of ten practice trials, arrows were replaced by the names of infections, randomly drawn from the pool of infections (and replaced for the main choice task). Responses from those practice trials were not included in the analysis.

Results

Before turning to Prediction 2, I describe the obtained inferences and recognition judgments in more detail. On average, participants scored 58.8% correct (*SD* = 4.9; *Mdn* = 59.2, range: 46.4% to 69.2%). The cap on response time resulted in somewhat fewer accurate choices, as a comparison with the average score in Study 1 shows ($t(98) = 1.99$, $p = .049$; $d = .41$). On average, participants recognized 61.0% (*SD* = 12.6, range 41.6%-100%) of the infections (Table 2.1.1). The recognition heuristic was applicable in 46.4% (*SD* = 11.8) of the pairs. A student of veterinary medicine recognized all 24 infections, thus rendering the application of the heuristic impossible. Therefore, the recognition validity α was calculated for only 59 participants. The average α was .62 (*SD* = .10), echoing the value obtained in Study 1 (.60). The knowledge validity β, however, was substantially lower than in Study 1: *M* = .62 vs. .66 ($t[98] = -3.18$, $p = .002$). It appears that under time pressure participants' ability to retrieve additional knowledge was compromised, thus giving way to more guessing responses when both infections were recognized.

Did time pressure increase adherence to the recognition heuristic (Prediction 2)? Consistent with Prediction 2, the proportion of choices in accordance with the recognition heuristic rose under time pressure. The mean proportion of inferences agreeing with the heuristic was 69.2% (*SD* = 10.7, range 41.4%-90.0%), compared to 62.1% in Study 1 ($t[63.5]$

= 2.5, p = 0.02, d = 0.55). Moreover, the variance in adherence rate (across participants) was smaller in Study 2 than in Study 1 (F[1, 97] = 10.6, p = .02). Note that this increase in the use of the recognition heuristic is not trivial. Time pressure could also have given way to mere guessing. In that case, the proportion of inferences agreeing with the heuristic would have dropped rather than risen. Instead it appears as if time pressure both fostered the use of the recognition heuristic *and* preempted the retrieval of more knowledge (thus attenuating β).

As Figure 2.1.7 shows, the increase in adherence to the recognition heuristic was also manifest on the level of individual infections. For 16 of the 23 infections (70%; as in Study 1, tularemia was not included), more choices agreed with the recognition heuristic than in Study 1 (also Table 2.1.1). In addition, that fact that five of the six diseases for which adherence dropped were among the seven diseases with the highest adherence rate in Study 1, thus suggesting a regression effect.

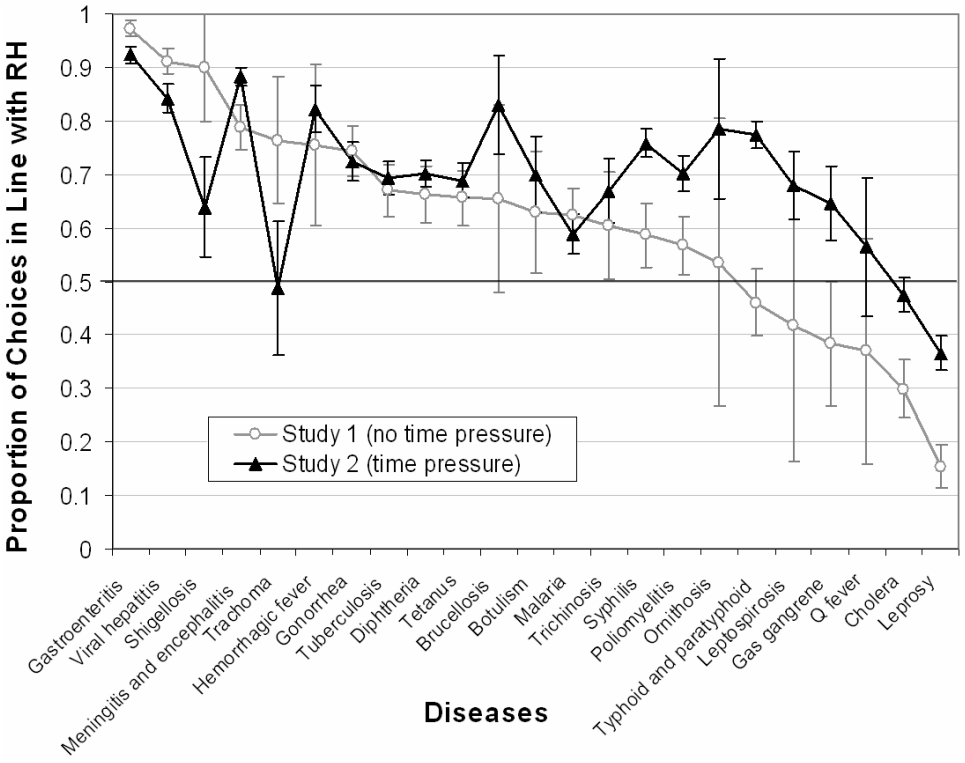

Figure 2.1.7. Object-specific adherence to the recognition heuristic for Studies 1 and 2 (cf. Figure 2.1.5). Tularemia, recognized by only one participant in both studies each, is not shown. Error bars indicate standard errors.

Were inferences in accordance with the recognition heuristic made faster (Prediction 1)? Study 2 also provides another test of Prediction 1. Specifically, it is possible to examine whether within the limited response-time window the inferences agreeing with the recognition heuristic decline as a function of time. Such an outcome would support Prediction 1 according to which inferences in line with the recognition heuristic are made faster than those that

conflict with it. I divided the response-time window into eight bins, starting with 400 ms to 499 ms, and ending with responses that lasted longer than 1,100 ms.[22] I then analyzed, for each bin and each infection (for which there were at least 100 choices within each bin; note that again the choices rather than the participants were taken as the unit of analysis), the proportion of choices in accordance with the recognition heuristic. Figure 2.1.8 shows the mean proportions in line with the recognition heuristic as a function of response time. Consistent with Prediction 1, proportions were above 70% for the early bins (i.e., 400 ms to 700 ms bins). For later bins, however, the mean proportion dropped rapidly. Consistent with Prediction 1, the more time a response took, the less likely it was consistent with the recognition heuristic (Prediction 1).

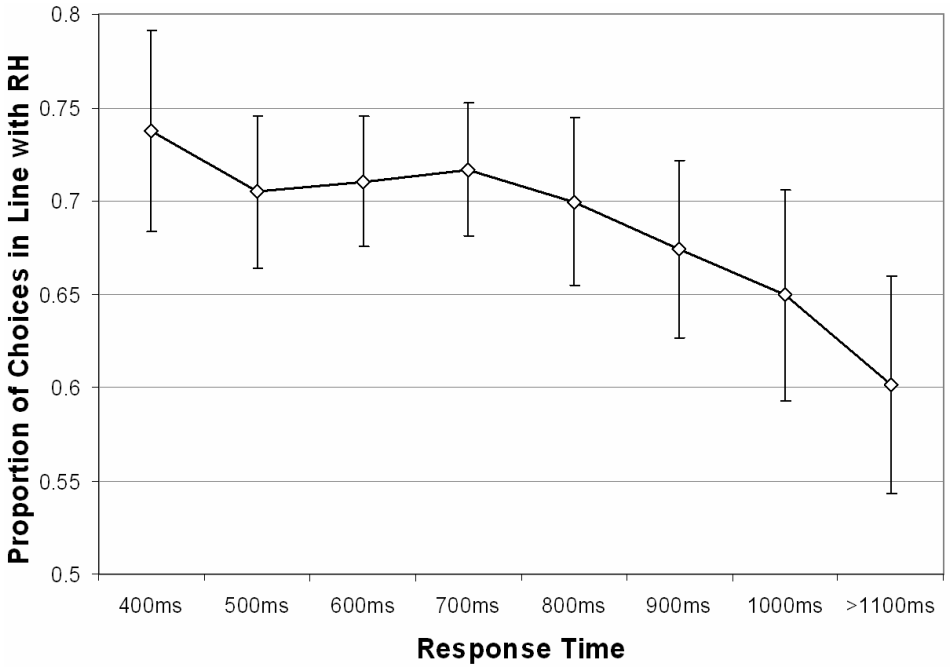

Figure 2.1.8. Proportion of choices following the recognition heuristic as a function of processing time. Over time, there is a decrease in the proportion of choices in line with the heuristic. The 15 diseases included were gastroenteritis, viral hepatitis, tuberculosis, meningitis and encephalitis, gonorrhea, syphilis, malaria, typhoid and paratyphoid, gas gangrene, hemorrhagic fever, tetanus, leprosy, cholera, diphtheria, and poliomyelitis. The number of choices in the eight bins from "400" to "> 1,100" were 141, 747, 1,639, 2,076, 1,354, 752, 260, and 267, respectively.

Did conclusive and conflicting criterion knowledge trigger the heuristic's suspension?

As Figure 2.1.7 shows, choices involving leprosy and cholera (when recognized and paired with an unrecognized disease) resulted in the lowest proportion of recognition adherence in Studies 1 and 2. Why was that? One possibility is that people assumed the diseases to be the least frequent ones. If so, any other disease (even if not recognized) can be inferred to be more frequent than either of the two. Consistent with this view, I found that both infections

[22] Since few responses took less than 400 ms, they were omitted them from the analysis.

produced lower frequency estimates than any other infection (see right most column in Table 2.1.1): The median estimate of their annual incidences was 5.[23] In addition, both infections were those for which the highest proportions of participants (30% and 31.7%, respectively; see Table 2.1.1) indicated that they had direct knowledge of incidence rates. These findings suggest that direct and conclusive criterion knowledge for the recognized infection—for instance, knowing that it is virtually extinct—appears to trigger the suspension of the recognition heuristic.

To assess more generally how conclusive criterion knowledge impinges on the use of the recognition heuristic, I reanalyzed people's heuristic adherence. To this end, I focused on those cases in which recognition and criterion knowledge conflicted, specifically those in which the frequency estimate for the recognized infection was *conclusively* lower than that for the unrecognized infection. Criterion knowledge was treated as conclusively lower if (a) the estimate for recognized infections in a pair was lower than the estimate for the unrecognized one by at least two category bins (e.g., the recognized infection was assigned to frequency category "2", the unrecognized one to "4", corresponding to the frequency ranges "10-99," and "1,000-9,999", respectively)[24], and (b) the recognized infection for which a person indicated having direct criterion knowledge was assigned the lowest possible frequency category (i.e., "1") by that person. Both conditions may give rise to a local mental model (Gigerenzer, Hoffrage, & Kleinbölting, 1991), thus rendering the retrieval and reliance on probabilistic cues unnecessary. Across participants, 866 cases met the two criteria. The recognition heuristic adherence, averaged across participants for which there was at least one such case, was below chance level, namely, 45.7%, and markedly lower than the same participants' adherence in all cases in which recognition discriminated (68.4%; $t[47] = -6.8$, $p = .001$). In addition, the mean adherence rate when recognition and criterion knowledge converged was 86.2%.[25] These results suggest that conclusive knowledge of the infections' criterion values can indeed mediate use of the recognition heuristic.

This conclusion was corroborated in a reanalysis of adherence to the recognition heuristic in Study 1. For this analysis, I took advantage of the median frequency estimates obtained in Study 2. Specifically, I focused on those 167 critical pairs in Study 1 that contained one unrecognized and one recognized infection, and in which the frequency estimate for the recognized infection was lower than for the unrecognized one by at least two

[23] I computed these values by replacing each of the six frequency categories (see Method section) with the midpoints of each category. For instance, the first category ranging from 1 to 9 was replaced by the value 5. The last category ">100,000" was replaced by the value 200,000.

[24] Interestingly, participants did not consistently give extremely low frequency estimates for unrecognized diseases. The mean estimated frequency (based on the midpoints of each category) for unrecognized infections was 2378.0 ($SD = 5771.4$), which was significantly different from the lowest frequency estimate ($t[57] = 3.13$, $p = .003$).

[25] Criterion knowledge that converged with recognition was defined as cases in which the frequency estimate for a recognized disease was higher than the frequency estimate for the unrecognized infection by at least two category bins (e.g., the recognized infection was assigned to frequency category "4", the unrecognized one to "2").

category bins. The recognition heuristic adherence was 19.1% (again, across all participants where there was at least one critical pair), around 44 percentage points lower than the adherence (of the same participants) in all pairs in which the heuristic could be applied (62.5%; $t[31] = -8.9$, $p = .001$).

Summary of Study 2

Consistent with Prediction 2, the mean proportion of inferences in accordance with the recognition heuristic increased under time pressure. That is, the competitive edge that recognition information enjoys over other knowledge—its retrieval primacy—translates into more judgments in accordance with the heuristic when people were pressed for time. In addition, additional evidence in support of Prediction 1 was found: The longer it took to arrive at an inference, the lower the proportion of choices conflicting with the recognition heuristic. Finally, I observed that conclusive and conflicting criterion knowledge appears to be a key triggering condition for the suspension of the recognition heuristic.

General Discussion

When Goldstein and Gigerenzer (2002, p. 77) proposed the recognition heuristic, they used the term recognition for the distinction between the "novel and the previously experienced." That is, in many situations an initial sense of recognition (or lack thereof) suffices to make a determination. The exemplary fast and frugal recognition heuristic does not require additional information such as in which context one encountered the object or what other knowledge about the recognized object one can marshal. Moreover, Goldstein and Gigerenzer assumed that this kind of information gives rise to noncompensatory inferences: If one object is recognized and the other is not, then the inference can be locked in. Since search for further information is then terminated, no other—conflicting—information about the recognized object can reverse the judgment suggested by recognition; simply, because it is not retrieved. However, this thesis of the recognition heuristic as a strictly noncompensatory strategy has been challenged. Oppenheimer's (2003; Experiment 1) results, for instance, suggest that direct knowledge of the small criterion value of an object can reverse recognition-based judgments. Newell and Shanks' results (2004) suggest that recognition-based judgments are reversed when other cues are available that conflict with recognition and when their validity is known to exceed that of recognition. In their view, recognition is a cue as any other.

I aimed at demonstrating that recognition—though not strictly noncompensatory as Goldstein and Gigerenzer (2002) envisioned it—is not like any other cue. To this end, I linked research on the heuristic with research on recognition memory. One relevant finding in this literature is, for instance, Gronlund and Ratcliff's (1989). They observed that information necessary to reliably judge whether a pair of words has been previously presented becomes available after a mere 350 ms of processing time. Decisions for which additional associative

information is required, in contrast, take considerably longer (around 570 ms). Based on these and similar findings, I proposed that mere recognition is already available while other information is still waiting in the wings. It is retrieved with little to no cognitive effort, while other knowledge needs to be searched for and looked up. These properties represent what I have termed recognition's retrieval primacy. Based on this notion, I have derived three predictions, and the evidence I obtained supports them.

Specifically, I found in Studies 1 and 2 that inferences in accordance with the recognition heuristic were made faster than inferences conflicting with it. In addition, reliance on the recognition heuristic increased when inferences had to be made under time pressure. Finally, I observed that in an environment in which recognition and criterion were not strongly correlated, the recognition heuristic was not as frequently used as in environments in which a strong correlation existed. Although there are likely to be others (see below), one key factor that triggers the temporary suspension of the use of the heuristic, ceteris paribus, seems to the presence of certain and conclusive knowledge of the criterion. In what follows, I discuss the implications of the results that were obtained. I first turn to a short review of the conditions that trigger the use and nonuse of the heuristic.

Under What Conditions do People use the Recognition Heuristic?

Newell (2005) criticized the fast-and-frugal-heuristics program (Gigerenzer et al., 1999) for having failed to "establish … [the] boundary conditions on the adaptive toolbox framework. Without such conditions, it is impossible to evaluate the adequacy of the proposed models of the decision processes" (p. 13). There are now a number of studies on the use of some of the tools of the adaptive toolbox, in particular the Take The Best heuristic (e.g., Rieskamp & Hoffrage, 1999; Bröder, 2000; Bröder & Schiffer, 2003; Newell & Shanks, 2004; Newell, Weston, & Shanks, 2003) and the recognition heuristic. Thus, it is now possible to draw upon an—admittedly preliminary—list of conditions that foster and hamper the use of the recognition heuristic.

Recognition validity. The recognition heuristic is *useful* when there is a strong correlation—in either direction—between recognition and criterion. But attending rather than ignoring can prove helpful, as Goldstein and Gigerenzer (2002) suggested, in any domain in which the recognition validity is higher than chance ($\alpha > .5$). In the current domain of infectious diseases the recognition validity proved higher than chance ($\alpha = .60$ and .62 in Study 1 and 2, respectively), and the observed adherence rates, though modest, suggest that recognition was indeed not ignored. Nevertheless, the adherence rate Goldstein and Gigerenzer (2002) observed in the city-size environment (90% in their Study 1) was substantially higher than in the infection environment that was investigated here (62.1% and 69.2% % for Studies 1 and 2, respectively). Though Goldstein and Gigerenzer did not report the recognition validities of their participants directly, one can take advantage of the

recognition correlation[26] they calculated for their city-size environment (p. 86). Comparing their correlation with the ones obtained here, there was indeed a much stronger correlation between recognition and the criterion: r_s = .01 and .03, (our Studies 1 and 2, respectively) and r_s = .66 (Goldstein and Gigerenzer's analysis). One interpretation of these parallel differences in adherence and recognition validity between the environments is that adherence to the recognition heuristic is at least partly contingent on the recognition validity in a domain. Such a dependency may reflect the user's adaptive and ecologically rational use of heuristics (Gigerenzer et al., 1999). Indeed, should one expect an adaptive user of the recognition heuristic to rely on the heuristic to the same extent, irrespective of whether recognition validity is .51 or 1?

However, one needs to be cautious not to overstate the degree to which the use of the recognition heuristic may attuned to α. Although there are indications for such a contingency across domains, individuals' use of the heuristic is typically not, or only moderately conditioned on their αs (see Studies 1 and 2; Chapter 2.2). In addition, the adaptive use of the recognition heuristic will be constrained by the accessibility of other knowledge. That is, even in environments low in recognition validity, recognition may often simply be the only accessible information. In the study of forecasts of the outcome of a sporting event reported in Chapter 2.2, laypeople appeared to rely almost exclusively on recognition in spite of a medium α of .7, and the existence of more valid cues. However, the authors represented cues typically available only to experts (e.g., team and player rankings, recent performance, etc.).

Conflicting knowledge. One can think of at least three kinds of information that may conflict with choice determined by the recognition heuristic, and thus trigger its suspension: (a) probabilistic cues with validities larger than α (Newell & Shanks, 2004); (b) source knowledge (e.g., object, say Chernobyl, is recognized for reasons completely unrelated to its size; see Oppenheimer, 2003), and (c) conclusive criterion knowledge (see Study 2). Judging from the available empirical evidence, all three kinds of knowledge may trigger a temporary suspension of the recognition heuristic. This is, perhaps, least surprising in the case of conclusive criterion knowledge. The very point of heuristics such as the recognition heuristic is to infer, on the basis of probabilistic cues, an unknown criterion. If the criterion is known or can be deduced, however, probabilistic inferences will become superfluous (see also the notion of local mental models as proposed by Gigerenzer, et al., 1991).

The precise extent to which the first two kinds of knowledge will trigger the suspension of the recognition heuristic is not clear. One reason is that it is unknown how well the experimental set-ups used to study the recognition heuristic generalize to real-world use of recognition. For instance, in Newell and Shanks's (2004) condition of Study 1—in which a low recognition adherence was found (*RL* condition)—recognition was induced experimentally, it was lower in validity than any other cue, and people were fully aware of

[26] Expressed as the rank correlation between the number of participants recognizing an infection and its prevalence rate.

recognition's inferiority, and the availability of superior cues (as they had the chance to learn the cues' validities through feedback). Moreover, one can speculate that the use of recognition is more likely in situations in which source knowledge of recognition is rather diffuse (Johnson, Hastroudi, & Lindsay, 1993), thus suggestive of an unspecific, unbiased source, and of natural mediation of the criterion variable (see Figure 2.1.1).

Time pressure. Time pressure is conducive to noncompensatory processing (e.g., Dhar & Nowlis, 1999; Payne, Bettman & Johnson, 1993; Rieskamp & Hoffrage, 1999; Svenson, Edland, & Slovic, 1990; Zakay, 1985; see Edland & Svenson, 1993, for an overview). It was also found that a cap on response time increased adherence to the recognition heuristic. This result, however, does not simply echo the frequent observation that under time pressure people appear to pay increased attention to the more important attributes in a decision context (e.g., Ben Zur & Breznitz, 1981; Böckenholt & Kroeger, 1993; Kerstholt, 1995; Payne et al., 1988; Wallsten & Barton, 1982). In my view, the reason why people make more use of the recognition heuristic under a limited response-time budget is that the retrieval of recognition information precedes that of other pieces of information and requires little to no cognitive effort. Because of these properties, I also suspect that not only time pressure but also, for instance, attending to a second task—while performing the choice task—would increase adherence to the recognition heuristic.

Recognition Information: Like Any Other Cue?

Based on its evolutionary roots, Goldstein and Gigerenzer (2002) highlighted recognition as a primordial psychological mechanism and conjectured that the recognition heuristic is a noncompensatory strategy whose choices cannot be reversed by additional pieces of information. Newell and Shanks' (2004) and Oppenheimer's (2003) findings are at odds with this conjecture, as are some of those that were found here. But does this mean that recognition is treated like any other cue, a conclusion that Newell and Shanks' advocated? I believe that the latter conclusion risks throwing out the baby with the bath water. Due to its mnemonic properties, recognition is different from other cues: It represents immediate, insuppressible, and inexpensive information. Studies 1 and 2 demonstrate the implications of these properties for inferences based on recognition.

But it is not only retrieval primacy that distinguishes it from other cues. The recognition heuristic only applies if one object is recognized and the other is not. Consequently, further search *in memory* would typically yield additional information (if any) about the recognized object only, but not about the unrecognized one. Such an asymmetry distinguishes recognition from other cues, and may make further search comparatively less likely. One possible reason why this may be so is that cue values, in particular, continuous cue values are difficult to evaluate when no simultaneous reference point—naturally provided by the other object's value—is available (see Hsee, 1996; Hsee, Loewenstein, Blount, & Bazerman, 1999).

Another property that distinguishes recognition from other cues is that recognition might sometimes be the only available cue. Shepard's (1967) and Standing's (1973) extensive

investigations of the recognition memory give testimony to the ease with which recognition knowledge can be acquired and to its durability in memory. As a consequence, recognition knowledge might often be the sole ground on which objects can be distinguished.

To summarize, there is evidence that conflicts with the conjecture that a choice determined by recognition cannot be reversed. Yet, so I argue, there are good reasons to suspect that recognition is not a cue like any other. I agree, and so I suspect would Goldstein and Gigerenzer, with Newell and Shanks (2004) that "relying solely on recognition ... is not necessarily the best policy" (p. 934). But because of our perceptual and cognitive architecture recognition is the first reason with which one is confronted, and because of constraints such as time pressure, cognitive load and inevitable lack of knowledge it will sometimes be the only one.

Which Mechanism Suspends the Recognition Heuristic, and When Will Suspension be Successful?

I intentionally investigated the recognition heuristic in a real-world domain in which recognition was not strongly correlated with the criterion (Figure 2.1.1). This 'hostile' environment was chosen to increase the likelihood of inferences that differ from those determined by recognition and thus to be able to test the predictions that follow from the notion of retrieval primacy. In this domain, I found that people's adherence to the recognition heuristic was low, compared to a domain with a strong correlation between recognition and criterion. How exactly the suspension of the heuristic is implemented is currently unclear, but some candidate mechanisms can be excluded. Users do not appear to execute some kind of threshold strategy that demands reliance on the heuristic as long as α is above threshold, or suspension of it as long as α is below the threshold (Prediction 3a). Similarly, users do not seem to adjust their reliance on recognition to α directly, the matching mechanism (Prediction 3b).

At this point, the most promising candidate is the suspension strategy (Prediction 3c). When sufficient time and cognitive resources are available people appear to be going through an evaluative step in which they assess aspects such as the value that recognition has for the inference task at hand, the availability of conclusive criterion knowledge, and perhaps, the availability of source information. On this view, the use of the recognition heuristic could be understood akin to a two-stage process proposed in recent memory models. Such models involve a production stage followed by an evaluation stage once memory is accessed (e.g., Whittlesea & Leboe, 2000). Indeed, some recent results of an fMRI investigation of the recognition heuristic (Volz, Schubotz, Raab, Schooler, Gigerenzer & von Cramon, 2005) suggest that recognition knowledge feeding into the heuristic might be subjected to such an evaluative filter.

How successful is such an evaluation? One important insight from the present studies is that the decision to temporarily suspend the heuristic may not automatically increase the

ultimate inferential accuracy. To do so, the validity of the knowledge that comes into play when recognition is dismissed must exceed recognition validity (for this selected set of items). Only then does the user of the heuristic benefit from thinking twice. This raises two interesting issues: First, in domains with a strong correlation between recognition and criterion, it is difficult to top the recognition validity. Thus, a strong association in combination with the unfavorable odds of finding even more valid information may foster the noncompensatory use of the recognition strategy in such domains. Second and conversely, a weak association in combination with the better odds of finding more valid information may foster the temporary suspension of the recognition heuristic, a speculation consistent with the results obtained here.

Conclusion

The recognition heuristic is a prototype of a fast and frugal heuristic. It uses recognition, a capacity that evolution has shaped over millions of years and that allows organisms to benefit from both their knowledge and their ignorance. Although in the domain of knowledge inferences recognition may not be as strictly noncompensatorily used as it seems to be employed in other evolutionarily important domains (possibly because the costs associated with false rejections are less lethal), it is distinct from other cues. The recognition heuristic can work so well under circumstances of limited time because it hinges on a cognitive architecture in which a complex process—recognition—is performed in a split second and without demanding many cognitive resources.

2.2 Does Ignorance Increase Forecasting Accuracy? The Use and Usefulness of Recognition in Lay Predictions of Sports Events.

Panem et circenses—thus did the poet Juvenal describe the efficient formula that the Roman emperors used to keep the population peaceful. More than just a description of a political strategy of its time, Juvenal's comment also reflects the early importance of sports games in society. It was not only watching the gladiator games and chariot races in Circus Maximus that amused the Roman people. Betting on the winners of the games was an equally important component of the event (e.g., Weeber, 1998). In this chapter, I am interested in people's forecasts for the contemporary equivalents of the Roman games, sports such as basketball, American football, or soccer. Underlining the close link between sports and prediction even today, Koehler, Brenner, and Griffin (2002) point out that "in no other domain are predictions more ubiquitous—or arguably less important—than in sports" (p. 709).

Previous research on the forecasting of sports events has primarily investigated expert judgment (Boulier & Stekler, 2003; Cantinotti, Ladouceur, & Jacques, 2004; Forrest & Simmons, 2000; Heath & Gonzalez, 1995; Kaplan, 1980; Koehler, 1996; Ladouceur, Giroux, & Jacques, 1998; Vertinsky, Kanetkar, Vertinsky, & Wilson, 1986). Consequently, only little is known of how laypeople, who have neither extensive domain-specific knowledge nor experience with elaborate judgment strategies, make forecasts for such events. This apparent neglect is surprising, as millions of non-experts engage in betting activities for all kinds of sports events every week. It is estimated that annually up to $380 billion are spent on sports betting (Macy, 1999), and in 2003, almost every fifth adult American had engaged in legal or illegal sports betting (American Gaming Association, 2004). To be sure, many of these bettors follow their sports of interest regularly and passionately. Yet, in light of the sheer number of bets, there is certainly a substantial proportion of bettors who have only superficial sports knowledge and bet only occasionally and mainly for hedonistic reasons. How do these people place their bets? Arguably, one important determinant of their betting behavior is which team or sportsman they think is more likely to win (though other factors such as emotional attachment often also play a role). In this chapter, I investigate how laypeople make such forecasts regarding winning, and I take advantage of the European Soccer Championships 2004 (EURO 2004) to study this issue.

Specifically, I test one recently proposed model of lay judgment, the recognition heuristic (Goldstein & Gigerenzer, 2002), and compare it to other candidate mechanisms in describing laypeople's forecasts. Second, I look at one theoretically predicted consequence of the recognition heuristic, the less-is-more effect, which refers to the phenomenon that knowing less can lead to higher accuracy than knowing more. To learn more about when it occurs in real judgments, I investigate the less-is-more effect both between groups differing in

knowledge—and for this purpose forecasts were also obtained from a group of experts—and across laypeople with varying levels of knowledge. Moreover, I evaluate the predictive value of recognition as a cue against a number of direct indicators of team strength (rankings, recent performance, odds) and also critically examine the experts' forecasts in light of the performance of "naïve" statistical models that take advantage of these indicators.

This chapter is organized as follows. The first part provides a brief overview of previous research on forecasting sports events, noting that modelling of forecasting strategies has primarily focussed on experts. I argue that the generalizability of the models developed for experts to laypeople is limited and propose the recognition heuristic as explicitly suited to modelling lay predictions. After describing the less-is-more effect, I propose four alternative mechanisms for how laypeople make forecasts, followed by a definition of the benchmark cues I used to evaluate the lay mechanisms in terms of predictive strength. Finally, I report a study on forecasts of matches at the EURO 2004 in which I test the candidate mechanisms and investigate the less-is-more effect.

Previous Research on Forecasting of Sports Events

How well can people forecast the outcome of sports events? Replicating, by and large, the findings on expert prediction for other domains (for an overview see, for example, Hastie & Dawes, 2001), research on sports experts has shown that their forecasting abilities are not very impressive. Although generally better than chance (Cantinotti et al., 2004; Forrest & Simmons, 2000; Ladouceur et al., 1998), their predictions are often not more accurate than those of non-experts (Andersson, Ekman, & Edman, 2003; Andersson, Edman, & Ekman, 2005; Heit, Price, & Bower, 1994). Moreover, forecasts by experts have frequently been found to be not better (and occasionally worse) than "naïve" forecasting strategies that consider only one cue (e.g., rankings or home-field advantage; Boulier & Stekler, 2003; Forrest & Simmons, 2000). Finally, even with increasing experience, sports experts do not seem to improve their predictive accuracy (Lebovic & Sigelman, 2001).

To account for the conspicuously unremarkable accuracy of expert predictions, the processes underlying the judgments became the focus of attention, and in fact evidence for suboptimality was found. For instance, studies on forecasts by sports fans showed that wishful thinking contributes to a large degree to mis-prediction (Babad, 1987; Buckley & Sniezek, 1992; Kaplan, 1980). Capturing expert predictions more directly, Boulier and Stekler (2003) examined the value of expertise for forecasting soccer games and found that if experts have knowledge that goes beyond publicly available knowledge (such as power scores and home-field advantage), they do not seem to use it. Moreover, Forrest and Simmons (2000), regressing forecasts by sports editors for British soccer matches on publicly available information about the performance of the teams (e.g., number of goals scored and conceded, points won, rankings), concluded that experts weighted the information not in accordance with its actual predictive value. In a similar vein, the experts studied by Cantinotti et al. (2004) often reported basing their forecasts primarily on the most recent performance, which

is an unreliable cue (e.g., Wood, 1992). Finally, Boulier and Stekler (2003) found experts' forecasts to be inferior to a "bootstrapping model" (i.e., a model derived from the experts' previous forecasts). In sum, research on expert prediction of the outcome of sports events suggests that experts integrate multiple pieces of information to make a forecast, but that the selection and weighting of this information is far from optimal.

Whereas sports experts enjoy the informational advantage of knowing a lot about the sports actors, laypeople (by which I mean persons who have no extensive domain-specific knowledge and follow the events in the domain only occasionally) have only a fraction of this information available and are less involved in the domain. How do they make their forecasts? Can the findings for sports experts be generalized to laypeople? Unfortunately, the number of studies that attempted to model laypeople's forecasts for sports events is very limited (a rare exception is Heit et al., 1994). However, strategies of experts and non-experts have been compared in other domains (e.g., Ofir, 2000; for an overview, see Camerer & Johnson, 1991), and the assumption in these studies is that for both experts and non-experts the same general model holds: information is weighted according to its (perceived) validity and then added. I argue, in contrast, that the generalizability of the results on expert prediction to lay prediction is limited.[27]

First, as noted above, by definition laypeople do not possess the same knowledge base as experts (and will thus necessarily rely on different cues), and they seldom engage in extensive information search prior to a forecast. Second, due to lack of experience and involvement, laypeople probably use less complex strategies than experts.[28] A popular approach to studying the processes underlying judgments is to model judgments with multiple regression, which integrates a large amount of information in a compensatory way. This approach has also been used in studies on forecasting of sports events (e.g., Boulier & Stekler, 2003; Forrest & Simmons, 2000).[29] However, pointing out the natural limits of the human cognitive apparatus, various researchers have questioned the psychological plausibility of computationally demanding mechanisms such as multiple regression (Simon, 1956; Gigerenzer & Selten, 2001; Gigerenzer & Kurz, 2001). This latter point should especially apply to lay judges. (It might be less valid for professional forecasters, who are often trained to use complex strategies.)

As an alternative to complex judgment models, it has been argued that people often rely on simplifying cognitive short-cuts—*heuristics*—when making judgments under uncertainty or solving problems (Kahneman, Slovic, & Tversky, 1982; Newell & Simon, 1972; Gigerenzer, Todd, & the ABC Research Group, 1999). Heuristics take into account

[27] Note, however, that it has also been argued that experts are prone to the same biases as non-experts (e.g., Koehler et al., 2002), which might suggest that they are using similar mechanisms (but see Johnson, Rennie, & Wells, 1991).

[28] Although some studies suggest that if given the same information, non-experts can use this information similarly to experts (e.g., Heit et al., 1994; Kaplan, 1980).

[29] For a discussion of the evolution of linear regression as an "as-if" model to a model of the actual cognitive processes underlying judgment see Brehmer (1994; also Gigerenzer & Kurz, 2001).

only very little information and often only one cue (Hogarth, 2001). Whereas people's use of heuristics in everyday decision making has often been studied in contexts in which the heuristics lead to errors (see Krueger & Funder, 2004; Funder, 1987), Gigerenzer and colleagues have recently brought forward a research agenda that sets out to understand how humans can successfully judge an uncertain world by using simple mechanisms that are matched to the structure of the environment. Their *fast and frugal heuristics* are simple decision models that do without integration of information and make a decision after looking up only very few cues. The prototype of a fast and frugal heuristic, the recognition heuristic (Goldstein & Gigerenzer, 1999, 2002), can be viewed as particularly appropriate to model lay judgment, as it not only tolerates but even requires incomplete knowledge. Recently, the recognition heuristic has also been examined in the context of sports events (Andersson et al., 2003; Serwe & Frings, 2004; Ayton & Önkal, 2004). Recognizing its property of acknowledging the specifics of lay forecasting (i.e., limited knowledge and computational simplicity) and its being ideal for the task of forecasting the winner in a pair-wise contest, I will further test the merits of the recognition heuristic and its consequences in lay predictions of sports events. In contrast to earlier work on the recognition heuristic, I will also compare it to alternative models of lay prediction and more rigorously examine the less-is-more effect, which will be explained in the next section.

Predicting Sports Events by Exploiting Partial Ignorance: The Recognition Heuristic

When faced with the task to predict, as in the study reported here on the EURO 2004, the winner of a soccer match, what are possible ways to proceed? According to the recognition heuristic (Goldstein & Gigerenzer, 2002), if only one of the teams is recognized, it is predicted that the recognized one will win, without considering any further information. Obviously, the recognition heuristic will not always apply, nor will it always make a correct forecast. Rather, it relies on (a) partial ignorance to make a forecast in the first place and (b) *systematic* ignorance to make that forecast accurately. To illustrate, if all teams that compete with each other are recognized, the recognition heuristic will not make any prediction at all. Moreover, whether a team is recognized or not should be systematic (rather than random) in the sense that recognized teams should be more successful. As successful teams are more often talked about it is likely that the teams one has heard of will also be, overall, successful in the future, so reliance on this environmental structure will often lead to correct forecasts. In general, it is said that in such an environment the heuristic is *ecologically rational*.

Two parameters describe the overall performance when using the recognition heuristic, when, for instance, forecasting the winner when a number of teams compete against each other. The *recognition validity*, indexed by α, is calculated as the proportion of cases in which the recognized team is the successful one divided by the number of matches where one team is recognized but not the other. The proportion of correct decisions when both teams are recognized is the *knowledge validity*, indexed by β. When both teams are unrecognized, one

has to guess. Goldstein and Gigerenzer (2002) pointed to an interesting consequence of using the recognition heuristic. They showed that when $\alpha > \beta$, there can be a *less-is-more effect*, in the sense that recognizing many or all teams leads to fewer (or the equal number of) correct decisions than recognizing only half of the teams. This occurs because if more than half of the teams are recognized the applicability of the recognition heuristic decreases and the higher α can contribute less to the overall performance (see Goldstein & Gigerenzer, 2002, for details).

The prediction of future events provides an important test case for the recognition heuristic, as forecasts cannot be contaminated with partial knowledge of the criterion value.[30] Indeed, there is evidence that laypeople actually use the recognition heuristic for making forecasts of sports events (although the boundary conditions of its application in general are still under debate; Newell & Shanks, 2004; Oppenheimer, 2003; see also Chapter 2.1). For instance, in Serwe and Frings (2004), tennis laypeople made forecasts for the matches at the 2003 Wimbledon tournament and also indicated which players they had heard of before. In more than 90% of the matches in which a recognized player played against an unrecognized player, the participants made the forecast that the recognized one would win. Moreover, the recognized player was the actual winner almost 70% of the time, indicating that the recognition heuristic was ecologically rational in this environment. Another interesting finding by Serwe and Frings (2004) was that recognition fared astonishingly well compared to other, tennis-specific cues, such as the Association of Tennis Professionals (ATP) rankings and betting odds. I will perform a similar comparison in the current study. Moreover, both Ayton and Önkal (2004) and Andersson et al. (2003) tested the recognition heuristic to model forecasts of soccer matches and found support for it (although Andersson et al. could not test it directly, as they operationalized recognition as the amount of knowledge about the teams and their prediction task did not involve paired comparisons).

Going beyond what has been done in previous studies, not only will I test how well the recognition heuristic describes lay predictions in general, but I will compare its performance with alternative strategies. Moreover, I will also test the less-is-more effect. The less-is-more effect has been demonstrated both in the lab and in real-world judgments. Ayton and Önkal (2004) had both British and Turkish participants make forecasts for English soccer matches. Surprisingly, although the English participants probably knew much more about the English soccer teams than did the Turkish participants, both groups performed about the same. Similar results were also reported by Andersson et al. (2003), who compared expert and lay predictions for the matches at the 2002 Soccer World Championships and found both groups achieving the same accuracy. In a non-sports domain, Goldstein and Gigerenzer (2002) presented participants repeatedly over a period of four weeks with a judgment task in which they compared pairs of American cities in terms of their population size. By judging the same cities several times, participants "learned" the names of some cities they had not heard of

[30] In some previous studies (Oppenheimer, 2003; Goldstein & Gigerenzer, 2002), it was possible that some participants knew the criterion value of some of the objects (for instance, that Berlin has 3.2 million inhabitants), which might have diluted the evidence for people's use of the heuristic.

initially. Although by the fourth week the participants recognized more cities than in the first week, their performance dropped—less knowledge had been better. Finally, Goldstein and Gigerenzer also reported a study in which American students, judging a representative sample of both German and American cities, performed better on the German cities, although they recognized all the American but only half of the German cities.

In sum, the existing evidence supports the notion that people with limited sports-specific knowledge use their lack of recognition—that they have heard of only one of two competitors—to make forecasts of the outcomes of sports events. Furthermore, the less-is-more effect associated with the recognition heuristic has been demonstrated empirically between groups, over time, and between environments (or domains). However, it has not been demonstrated *across participants* varying continuously in level of knowledge, although Goldstein and Gigerenzer's (2002) computer simulation suggests it should. In the present study, I try to trace such a less-is-more effect across participants.

Study 3

Overview

Before the tournament took place, groups with different degrees of knowledge (laypeople and experts) were asked to forecast the winners of the 24 first-round matches of the EURO 2004 and to indicate for each participating country whether they had heard of the national soccer team of that country. The adherence to the recognition heuristic was tested by looking at how often the recognized teams, when playing again an unrecognized team, were judged to be more likely to win. The performance of the recognition heuristic in describing participants' forecasts was compared to alternative models of lay prediction (described below). To assess the predictive accuracy of the participants and the mechanisms tested, I compared the predictions with the actual outcomes of the matches.

To test the less-is-more effect, I used two approaches: First, I examined—between groups—if the less knowledgeable laypeople would make better forecasts than the knowledgeable experts. Second, I examined—within the group of laypeople—if individuals who recognized fewer teams would make better forecasts than individuals who recognized more teams.

Why should people rely on recognition to forecast the winners of the matches of the EURO 2004? Although a person lacking specific knowledge about the strength of the teams cannot directly make a prediction of how successful the teams will be, it can be assumed that strong teams have been mentioned in the media more frequently and further that more frequently mentioned teams are also be more likely to be recognized. In that sense the media act as mediators, relating team strength and recognition (see Goldstein & Gigerenzer, 2002). To test this assumed link between team strength and frequency of mentions in the media, I counted the co-occurrence of the country names with the term "soccer" in the German print

media (using the COSMAS II word corpus[31]) and correlated the number of co-occurrences with the Fédération Internationale de Football Association (FIFA) ranks of the teams, a plausible indicator of team strength. As predicted, there was a strong rank correlation between FIFA rank and mention frequency in the media, r_s = -.77. Furthermore, using the recognition rates collected in the present study (see below), the mention frequency and the proportion of lay participants recognizing the teams were highly correlated, r_s = 0.94. Moreover, recognition was correlated with FIFA ranks r_s = -.81, strongly supporting the assumed interconnectedness and illustrating the ecological rationality of the recognition heuristic for the forecasting task.

Alternative Mechanisms of Lay Prediction

One problem inherent in previous studies on the recognition heuristic is that as a model of people's judgments, the heuristic was not tested against alternative models—making it difficult to evaluate how well the heuristic describes what people actually do. It is possible that people are not using recognition but instead look up another cue, and if recognition is correlated with this other cue, the heuristic could still describe the data well. Therefore, here I contrast the recognition heuristic with four alternative accounts of how people with limited soccer knowledge forecast the winner of the matches at the EURO 2004.

What are other plausible cues that laypeople might use to forecast the winner of a soccer match with national teams? Although laypeople have by definition only little soccer-specific knowledge about the teams (e.g., their recent performance, current players), they can be expected to know something about the participating countries (e.g., population size, economic strength). Such general knowledge is often used in real-world estimation task (Brown & Siegler, 1993), and since it might also be probabilistically related to success in sports, this knowledge could also be used to judge which team is more likely to win. In the following, I describe three alternative cues that could correlate with team strength, on which laypeople might base their forecasts. In addition, I describe an additive model that integrates these three cues.

Gross domestic product (GDP). The rationale of this strategy is that richer countries can be expected to invest more money in the promotion of talented soccer players than poorer countries. Therefore, this strategy predicts that the team from the country with the higher per capita gross domestic product (as an indicator of economic strength of the country) is judged to be more likely to win.

Population size of the country (POP). All things being equal, countries with a large population will produce a larger number of outstanding players than countries with a small

[31] COSMAS (Corpus Search, Management and Analysis System) is the largest online archive of German literature (e.g., encyclopedias, books, and newspaper articles) and can be accessed on http://www.ids-mannheim.de/cosmas2/. At the time of the analysis, the corpus contained a total of the around 1.9 billion words. For the search, a co-occurrence was defined as the country name and the term "soccer" occurring together within a maximum span of ten words.

population. According to this strategy, the country with the larger population is judged to be more likely to win.

Former membership to East Block (EAST–WEST). A distinctive characteristic differentiating European countries is whether they belonged to the former East Block. Using this feature, EAST–WEST predicts that when a Western European team plays against a team from a country that belonged to the former East Block, the Western European team is judged to be more likely to win. EAST–WEST has a number of interesting properties. As most formerly East Block countries have only relatively recently joined the European Community, many people in Germany, where Study 3 was run, will not have heard of the national soccer teams of these countries. As a consequence, the predictions of EAST–WEST will show a large overlap with the predictions of the recognition heuristic. Furthermore, many of the East Block countries are still relatively weak economically and have relatively small populations (the exception being Russia). Therefore, EAST–WEST also implicitly combines some of the cues of the other mechanisms described above.

When testing GDP and POP it was not expected that participants would have precise knowledge of the relevant variables. Rather, I assumed a difference threshold δ, and the values a and b of two countries were perceived as different only if they differed by at least 20% of the larger value ($\delta = \max\{a,b\} \times 0.2$). Otherwise, the mechanisms do not discriminate, and no prediction is made.

That people rely exclusively on one cue when making a forecast is certainly a bold claim. Therefore, I also tested a model that integrates the three cues: GDP, POP, and EAST-WEST.

Tallying (TALLY). According to this model, the (equally weighted) positive evidence for each of the teams is added up and the team that has a larger sum of positive evidence is selected to be the one more likely to win (e.g., Dawes, 1979). Note that this model does not weigh the cues differentially but simply counts the positive and negative aspects for each team, which makes the model psychologically plausible. TALLY integrates three cues: population size, GDP, and EAST–WEST; thus a maximum sum of 3 is possible.

Apart from determining how well the proposed simple mechanisms describe what laypeople do, I was also interested in how well the mechanism that accounts for the forecasts can predict the winning team in a soccer game. To evaluate this, a comparison to the performance of established benchmarks is required. In the next section, I describe four such benchmarks.

Setting the Benchmark for the Mechanisms of Lay Prediction: Direct Indicators of Team Strength

Previous studies have identified rankings (Boulier & Stekler, 1999; Serwe & Frings, 2004), past performance (Boulier & Stekler, 2003; Forrest & Simmons, 2000), and betting odds (Boulier & Stekler, 2003; Serwe & Frings, 2004) as robust predictors of success in

sports. Aiming for the strongest possible test, I chose these direct indicators of team strength to obtain benchmark levels of performance against which the mechanism describing lay participants' forecasts best is compared. In addition, I also used the accuracy of the experts *as a group* to obtain another benchmark. According to the Condorcet jury theorem (Condorcet, 1785/1994), the pooled judgments of a group of individuals can be more accurate than that of the average individual (Grofman & Owen, 1986), and Forrest and Simmons (2003) showed such an effect in the sports domain as well.

FIFA ranking. The FIFA rankings constitute an amalgam of a large amount of information concerning the performance of the teams in the last eight years such as wins, number of goals, whether the match was at home or away, importance of the match, and regional strength (with less recent performance given progressively less weighting). I tested how often the team with the higher FIFA rank before the tournament actually won.

Performance in qualifying round. Teams were ranked in accordance with their performance in the qualifying round for the championships (there were no ties). Relevant criteria were points won (3 points for a win, 1 point for a draw) and the goals scored and conceded. (Because Portugal, as host team, qualified automatically, the matches that included Portugal were not included in the test of this cue.) I tested how often the team with the higher rank according to performance in the qualifying round won.

Betting odds. The betting odds were taken from online bettors[32]. It was tested how often the outcome with the lower odds (odds for draws were excluded) actually occurred. It should be noted that for two reasons, the strength of the odds for predicting success should not be too surprising: First, odds represent an aggregate of predictions by a large number of people. Second, in contrast to the other indicators, odds are continually updated, so they contain information about the course of the tournament.

Expert majority. For every match, I counted how many of the experts in Study 3 had selected each team to be more likely to win, and I checked how often the team that the majority had selected actually won. One match (The Netherlands vs. Czech Republic) was not included in the analysis as an equal number of experts had selected each team.

Hypotheses and Research Questions

Before I move on to the study, here is an overview of the hypotheses for Study 3:

- The recognition heuristic will describe laypeople's forecasts better than the other candidate mechanisms.
- If $\alpha > \beta$ (i.e., if the recognition validity is higher than the knowledge validity), and experts and laypeople have comparable values on these parameters, there will be a less-is-more effect between groups: Experts will not make more correct forecasts than laypeople.

[32] Odds were obtained from BetExplorer (http://www.betexplorer.com/soccer/international/euro-2004/league.php?group=0&lastXMatches=9999&round=1), which provides a summary of the final offers of around 70 online betting companies.

- If $\alpha > \beta$, there will be a less-is-more effect across lay participants: In other words, at some level of $n < N$ an increased amount of knowledge is not associated with more accurate forecasts.

In addition, I sought to obtain an answer to the following research questions:

- How well does the mechanism that describes the lay forecasts predict the actual winners of the matches compared to direct indicators of team strength, such as rankings, recent performance, and betting odds?
- How accurate are the experts' forecasts compared to "naïve" statistical models that are based on a single direct indicator of team strength?

Method

Participants

Laypeople. One hundred and twenty-one (62 women and 59 men; mean age 29.9 years, range 11-72) participants were recruited at various public places in Berlin (cafeterias, museums).

Experts. To recruit soccer experts, editors of sports sections at major German newspapers as well as TV and radio stations were contacted via email. They were asked for participation in the study and to forward the announcement to colleagues. Twenty soccer experts (2 women and 18 men, mean age 39.5 years, range 28-65) participated.

Tasks and Measures

Interest and knowledge. To check whether the recruited individuals were indeed from the desired populations, participants were asked to indicate their interest in soccer on a 7-point scale, ranging in interest from 1 (not at all) to 7 (very much). They also rated their soccer knowledge on a 7-point scale, ranging from 1 (much worse than average) to 4 (average) to 7 (much better than average).

Forecasting task. Participants were presented with a randomly ordered list of the 24 matches of the first round of the tournament (both the positions of the teams and the order of the matches were determined randomly) and asked to indicate which team they thought was more likely to win the match.

Ranking task. Participants were given an alphabetically ordered list with the names of the participating countries and asked to give ranks to each of the 16 national teams such that the rank given to a team would denote that the team was also more likely to win against all teams with a lower rank. I used the rankings to test the robustness of some of the results of the predictions for the first round (see Footnote 34).

Recognition task. Participants were presented with the names of the countries in alphabetical order and asked to indicate whether they had heard or read about the national soccer team of each country.

Procedure

The data collection took place in early June 2004, within the 14 days preceding the EURO 2004. The questionnaires were administered to the laypeople as a paper and pencil task, whereas the experts completed an online version. After the participants had indicated their age, sex, and profession, they completed the tasks in the above order. All participants were offered a monetary incentive for the forecasting task: They were informed that the participants with first, second, and third best accuracy would receive 30€, 20€, and 10€, respectively. Whereas experts were assumed to be intrinsically motivated to participate, laypeople received a chocolate bar as compensation. The completion of the questionnaire for laypeople took around 15 minutes (I have no reason to assume that the experts had different completion times).

Results

One lay participant gave extremely irregular responses (more then two standard deviations on crucial variables such as accuracy, α, and β) and was thus excluded from the analysis.

Knowledge and Interest

On the 7-point scale, the lay participants reported, on average, an interest in soccer of 3.02 (*SD* = 1.87) and a level of soccer knowledge of 2.79 (*SD* = 1.74). Note that they rated both their knowledge below the level marked "average". The experts, by comparison, reported an average interest in soccer of 6.8 (*SD* = 0.4), and an average knowledge of 6.1 (*SD* = 0.8). Consistent with their professional status as soccer experts, experts reported higher levels for both interest (Cohen's *d* = 1.88) and knowledge (*d* = 2.31). Next I turn to the question of how well the recognition heuristic and, by comparison, the other candidate mechanisms describe lay participants' forecasts.

Could Participants Use the Recognition Heuristic?

On average, the lay participants recognized 11.1 (69.2%) of the 16 teams (*SD* = 3.4; range 3-16). Overall, they were able to use the recognition heuristic for, on average, 9.22 (*SD* = 4.93) of the 24 matches (38.4%), yielding an average discrimination rate (DR) of the recognition heuristic of .38. The experts, in contrast, recognized, on average, 15.7 (97.9%) of the 16 teams (*SD* = 1.1). Eighteen of the 20 experts could never use the recognition heuristic and 2 could use it 7 and 10 times, respectively. (Overall, the recognition heuristic was thus applicable for the experts in only 0.85 (*SD* = 2.66) of the 24 matches.)

Forecasts in Line with the Recognition Heuristic

I looked at all matches where a participant had heard of one of the teams but not the other, irrespective of the outcome of the match. I then checked how often the recognized team

was selected to be more likely to win. Across the 103 lay participants who could apply the recognition heuristic at least once, when one team was recognized but not the other, the mean proportion of forecasts in line with the recognition heuristic was 90.5% (*SD* = 12.2; median and mode were 100%; see Table 2.2.1). As noted above, only two expert participants had the necessary ignorance to be able to use the recognition heuristic. (These two experts selected the recognized team in, on average, 70.7% (*SD* = 1.01) of the cases.)

How Well Did the Other Candidate Mechanisms Describe Laypeople's Forecasts?

To test how well the alternative accounts, GDP, POP, EAST–WEST, and TALLY, described which team the lay participants selected to be more likely to win, I determined for each mechanism how often, when it discriminated, its predicted forecast coincided with the forecast of the participants. Figure 2.2.1 and the upper part of Table 2.2.2 show how well the different mechanisms described the lay participants' forecasts, along with the discrimination rate of each mechanism (i.e., the proportion of matches where the mechanism made a prediction). To facilitate a direct comparison with the recognition heuristic, in Figure 2.2.1 only those matches where the recognition heuristic made a prediction are considered. In the lower part of Table 2.2.2 the overall proportions of correct predictions of the different models are reported. (The results are essentially the same.)

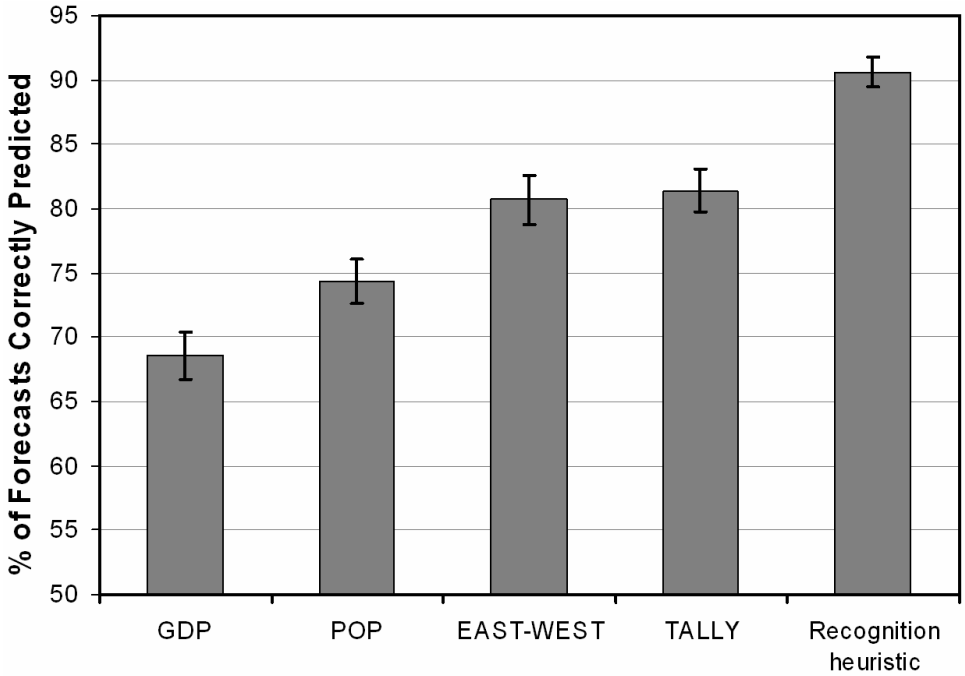

Figure 2.2.1. Mean proportion (across participants) of lay participants' forecasts correctly predicted by the five mechanisms tested. See text for description of the mechanisms.

Table 2.2.1. *Results for the first-round matches. RH Recognition heuristic.*

	Laypeople	Experts
Accuracy (% correct)	64.5[a]	76.6[a]
SD	10.6	9.3
Percentage of forecasts in line with RH	90.5	70.7[b]
SD	12.2	25.3
Recognition validity α	.71[a]	.68[a,b]
SD	.18	.25
Knowledge validity β	.60[a]	.77[a]
SD	.24	.10

[a] The eight matches that ended in a draw are not included in the analysis.
[b] Results based only on $n = 2$.

Table 2.2.2. *How well do the five different mechanisms of lay prediction describe lay participants' forecasts? The upper part shows the proportion of forecasts that were in line with each of the five mechanisms for those cases in which a participant recognized one team but not the other. The lower part shows the results when all matches where the individual mechanisms discriminated are considered. See text for description of mechanisms. DR discrimination rate.*

		POP	GDP	EAST-WEST	TALLY	Recognition heuristic
Only matches where individual recognition discriminated	Mean proportion of correctly predicted forecasts	.74	.69	.81	.81	.91
	SD	.18	.19	.20	.17	.12
	DR	.35	.35	.24	.29	.38
All matches where cue discriminated	Mean proportion of correctly predicted forecasts	.68	.62	.75	.73	.91
	SD	.09	.10	.14	.11	.12
	DR	.88	.83	.54	.75	.38

Of the four alternative mechanisms, TALLY described the laypeople's forecasts best, but it still laged almost 10 percentage points behind the recognition heuristic. This difference was even greater when all matches of the first round are considered, though the mechanisms differed considerably in terms of their DR.

A possible objection to the preceding analysis is that the recognition heuristic had an advantage due to its predictions being based on individual data. It could, thus, for a given match, make different predictions for different participants, whereas the other mechanisms made the same prediction for each individual. To deal with this objection, I also tested a version of the recognition heuristic that made the same prediction for each participant. I used

collective recognition (see Goldstein & Gigerenzer, 2002) for this purpose, and thus it was predicted that the team that more participants had heard of would be selected (to ensure comparability with the other mechanisms, again a threshold of 20% of the larger value was assumed). It turned out that even with this stricter test, on average 87.3% (*SD* = 11.5) of lay participants' forecasts were in line with the collective recognition heuristic (DR = 0.75). Thus, the considerably higher proportion of predictions in line with the recognition heuristic is not due to its predictions being tailored to the individual participants.

To conclude, of five plausible models of how people without soccer-specific knowledge make forecasts for the soccer matches, the recognition heuristic described the data for the lay participants best. More than half of the participants who had not heard of all teams *always* selected the team predicted by the recognition heuristic when possible.

To understand further why the recognition heuristic was preferred as a forecasting strategy, I calculated how well the five candidate mechanisms predicted the actual outcomes of the matches. Eight of the 24 matches of the first round ended in draws and these matches were not included in the accuracy analysis as a draw was not modeled by the mechanisms (nor was a draw a response option in the participants' forecasting task). Table 2.2.3 reports the proportion of correct predictions (i.e., the validities) and how often each mechanism discriminated. All mechanisms predicted better than chance (0.5), and the recognition heuristic turned out to have the highest validity, but also the lowest DR.

The previous analysis established the recognition heuristic as an appropriate model to describe the forecasts of the lay participants. I now turn to the question of whether using the recognition heuristic led to participants who had heard of fewer teams making more correct forecasts than participants who had heard of many or all of the teams.

Table 2.2.3. *Ecological validities (i.e., proportion of correct predictions of actual results) and discrimination rates (DR) of the lay mechanisms (only the 16 non-draw matches are considered). As individual recognition varied across participants, the mean ecological validity of recognition (α) is reported along with the standard deviation.*

		POP	GDP	EAST-WEST	TALLY	Recognition heuristic (*α*)
All matches where cue discriminated	*Ecological validity*	.54	.57	.70	.67	.71
	SD					.18
	DR	.81	.88	.63	.75	.39

Was There a Less-Is-More Effect?

I was interested in two possible less-is-more effects, between groups (experts vs. laypeople) and across laypeople spanning different levels of knowledge. In the following, I analyse the accuracy of the participants as a function of their level of knowledge, that is, the number of teams they had heard of. Again the matches that ended in draws were not included in the accuracy analysis.

Ecological analysis. Recall that Goldstein and Gigerenzer (2002) demonstrated that in order for a less-is-more effect to occur, the recognition validity α must be larger than the knowledge validity β. To check whether this condition held, I determined α and β for the lay participants and for the experts separately. The results of the ecological analysis are shown in Table 2.2.1.[33] For the lay participants, the conditions for a less-is-more effect were met: the average α (.71) was larger than average β (.60). Keeping in mind that α for the experts was calculated only for the two experts who had not heard of all the teams, it is interesting that the experts' α was similar to laypeople's α. Importantly, however, the experts had a considerably higher β than the laypeople, $t(68.2) = 3.01$, $p = .001$, $d = .73$, thus one should not expect the experts to perform worse than the laypeople. Given these results, a less-is-more effect is predicted within the group of laypeople (when forecasting accuracy is analysed as a function of the number of recognized teams), but not between experts and laypeople. Accordingly, for the 16 non-draw matches, the laypeople made on average 64.7% correct forecasts (*SD* = 10.3; range 37.5-85.5%), whereas the experts predicted significantly better and made on average 76.6% (*SD* = 9.27; range 56.3-93.8) correct forecasts, $t(138) = 4.8$, $p = .001$, $d = 1.16$.[34]

Less-Is-More Across Lay Participants. Using the average values of α (.71) and β (.60) of the lay participants and Equation 1 in Goldstein and Gigerenzer (2002, p. 78; see Chapter 1.4.1), I calculated the expected overall accuracy for each level of *n*. Figure 2.2.2 shows the predicted pattern. The highest accuracy (64.3% correct forecasts) is predicted at $n = 10$. For *n* higher than 10, the predicted accuracy decreases, reaching 60% when all teams are recognized. Less should be more.

Could such a pattern be found in the data? Figure 2.2.3 shows the observed and predicted accuracy as a function of *n*, for the 83 participants with at least 90% of their forecasts in line with the recognition heuristic (although lay participants who recognized all 16 teams could never use the recognition heuristic, they were nevertheless included in this analysis as they provided an important reference point: complete knowledge). For the number of recognized teams, the data were collapsed in five bins. (Due to low cell frequencies, cases with eight or fewer recognized teams were collapsed in the first bin. The lowest number of recognized teams was $n = 3$.) As a first observation, the overall accuracy was higher than predicted (with the exception being the bin with $n = 9$-10). Second and importantly, across the different levels of *n*, the average proportion of correct forecasts increased until $n = 13$-14, but

[33] For the calculation of the recognition validity and knowledge validity, only those matches that were not draws were considered. Note that both recognition validity and knowledge validity could not be calculated based on all possible pairs of teams, but only on the matches actually played. From the matches played, only the matches of the first round were used, as the sample of possible matches played in the first round is most representative here. The matches in the knock-out phase represent a much more selected sample, as only strong teams are included, and including these matches would possibly distort the resulting indices.

[34] Similar results also held for the matches of the knock-out phase (seven matches in total). Predictions for these matches were derived from the participants' rankings (see Method section). The mean consistency between predictions and rankings for the first round were, on average, 83.9% (*SD* = 12.2) for laypeople and 87.9% (*SD* = 7.8) for experts. Though both experts and laypeople predicted much worse than for the first round, the experts (47.9% correct, *SD* = 10.7) still predicted considerably better than the laypeople (39.3% correct, *SD* = 15.6), $t(139) = 2.36$, $p = .02$.

with n = 15-16, the proportion of correct prediction dropped—the predicted less-is-more pattern. In other words, across the laypeople at the highest level of knowledge recognizing more teams does not mean higher accuracy.

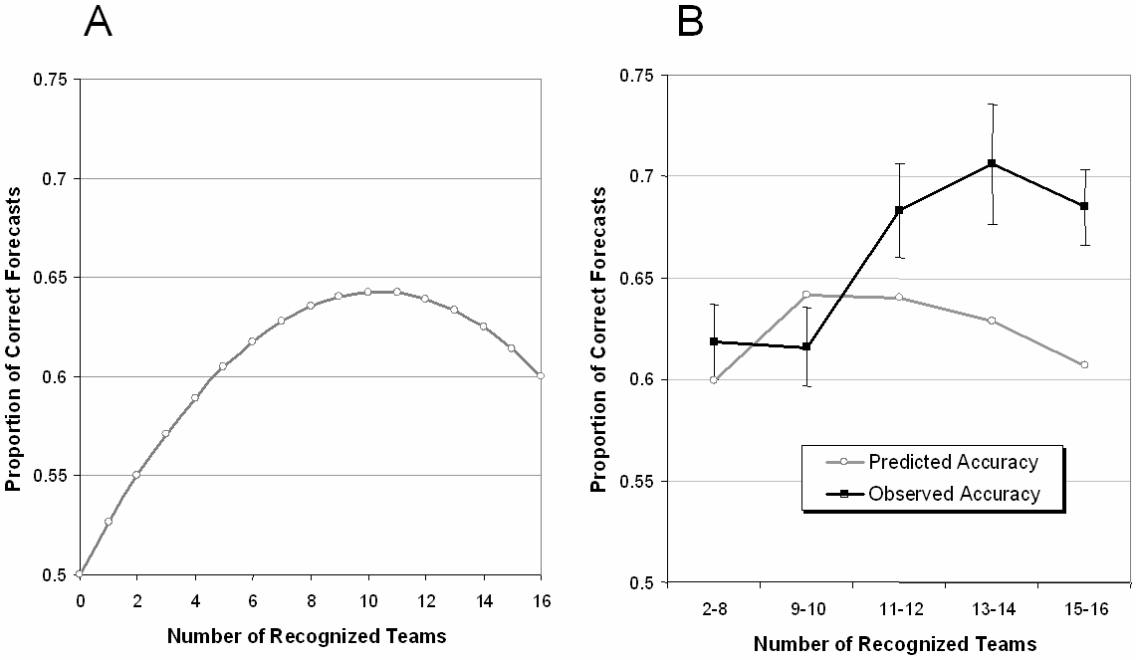

Figure 2.2.2. (a) Predicted forecasting accuracy as a function of n (using Equation 1 of Goldstein and Gigerenzer, 2002), based on the average observed α (.71) and β (.60). (b) Less is more: Predicted and observed forecasting accuracy as a function of n. Five bins were created in incremental steps of n = 2 (with exception being the first bin; see text for details). For the observed accuracy, the number of participants in the five bins was 19, 15, 14, 10, and 25, respectively. As for the whole sample of laypeople, for this subgroup the average α (.72) was larger than the average β (.61). The bars show standard errors. The predicted accuracies for each bin were determined by averaging the values for the different levels of n contained in each respective bin.

It should be noted, however, that compared to the less-is-more effect expected from the calculation based on Goldstein and Gigerenzer's Equation 1 (shown in Figures 2.2.2a and b), the benefit of less knowledge emerges at a higher level of knowledge than predicted. The predicted accuracy depicted in Figure 2.2.2a peaks at n = 10, whereas the observed maximum accuracy in Figure 2.2.2b is shifted to the right and peaks at n = 13-14.

What is the reason for this shift? A look at the relationship between the number of recognized teams, n, and both α and β reveals that α and β were both *positively correlated* with n, $r_{\alpha n}$ = .19 (p = .05), and $r_{\beta n}$ = .22 (p = .02). In Goldstein and Gigerenzer's simulations, in contrast, n was assumed to be unrelated to α and β. What does this correlation imply? As α and β are the main determinants of accuracy, if both α and β increase with increasing n there will be a positive relation between n and overall performance—the opposite trend of a less-is-more pattern. If it is true that the correlation of n with both α and β can shift the less-is-more

effect, the turning point at which more knowledge is associated with less accuracy should appear earlier when α and β are kept constant.

To test this, participants with similar values of α and β were grouped together and their forecasting accuracy plotted as a function of *n*. Because minimal conditions for the less-is-more effect are use of the recognition heuristic and that α > β, only participants who applied the recognition heuristic in at least 90% of the possible cases and for which α was larger than β were included in the analysis. The groups were constructed as follows. First, for both α and β, the range from 0.5 (chance level) to 1 was divided into three sub-ranges (0.5-0.66, 0.67-0.83, 0.84-1) and participants who were in the same sub-range for α and β were pooled in one group. Five of the nine possible groups had a cell frequency larger then zero. Figure 2.2.3 depicts these five groups. As visual inspection reveals, whereas in the left half of Figure 2.2.3 the direction of the lines is generally upward, in the right part of the figure the lines are pointing downward: the less-is-more effect now emerges at intermediate levels of knowledge. In four of the five groups that met the inclusion criteria, participants who recognized the most teams had a lower (or equal) proportion of correct predictions than participants who recognized only about half of the teams.

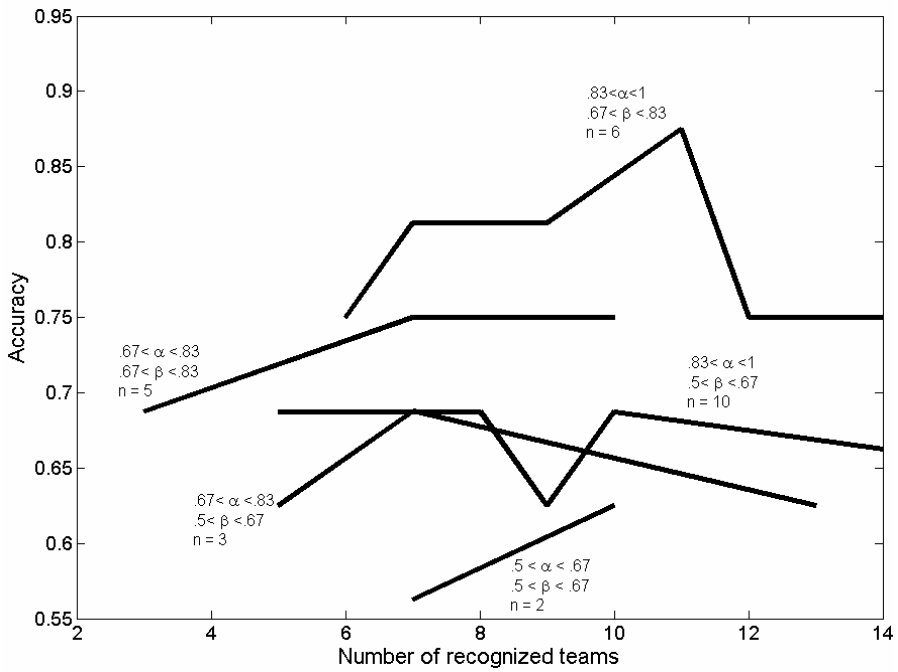

Figure 2.2.3. Less is more holding α and β constant.

In the previous sections it was found that of the candidate mechanisms tested, the recognition heuristic describes the forecasts of the lay participants best and one consequence of using the recognition heuristic was examined, the less-is-more effect. Next, I turn to the third main issue considered in this chapter and evaluate how well recognition predicts the winners at the EURO 2004 tournament compared to direct indicators of the strength of the

teams (which I assumed were unknown to the lay participants, but possibly known by the experts).

How Accurate Was Recognition Compared to Direct Indicators of Team Strength?

To evaluate the ecological rationality of the recognition heuristic in the environment studied, the prediction of the winners of the EURO 2004, I submitted recognition to a tough test and contrasted it with the predictive strength of direct indicators of team strength (rankings, past performance, betting, odds and expert majority). To obtain one single recognition measure and thus make recognition directly comparable to these indicators, which were continuous variables, I followed Goldstein and Gigerenzer (2002) in using collective recognition (i.e., the percentage of lay participants who recognized each team). Figure 2.2.4a shows the proportion of matches correctly predicted by collective recognition compared to the four direct indicators of team strength. The results of the analysis with individual, rather than collective recognition are shown in Figure 2.2.4b (based on only those matches where individual recognition discriminated, averaged across participants). The results are essentially the same: recognition, with 63% correct predictions, is clearly beaten by the competitors. The predictions based on the majority of experts achieved the highest accuracy (87%), closely followed by the teams' performance in the qualifying round (85%).

How Accurate Were Direct Indicators of Team Strength Predict Compared to Individual Experts?

Finally, Figure 2.2.4a also invites a critical analysis of the accuracy achieved by the experts in the present study. The experts were outperformed by two simple statistical models: both picking the team with a higher FIFA rank (81%), and picking the team with the better performance in the qualifying round (85%) would have resulted in higher performance than the experts' average individual forecasting accuracy (indicated by the upper dotted line; for comparison, the lower dotted line represents the average individual accuracy of the lay participants). Additionally, always selecting the team that the majority of the experts judged to be more likely to win would have resulted in 87% correct forecasts, considerably higher than the 76.6% achieved by the individual experts. Only 4 of the 20 individual experts achieved the same or a higher accuracy.

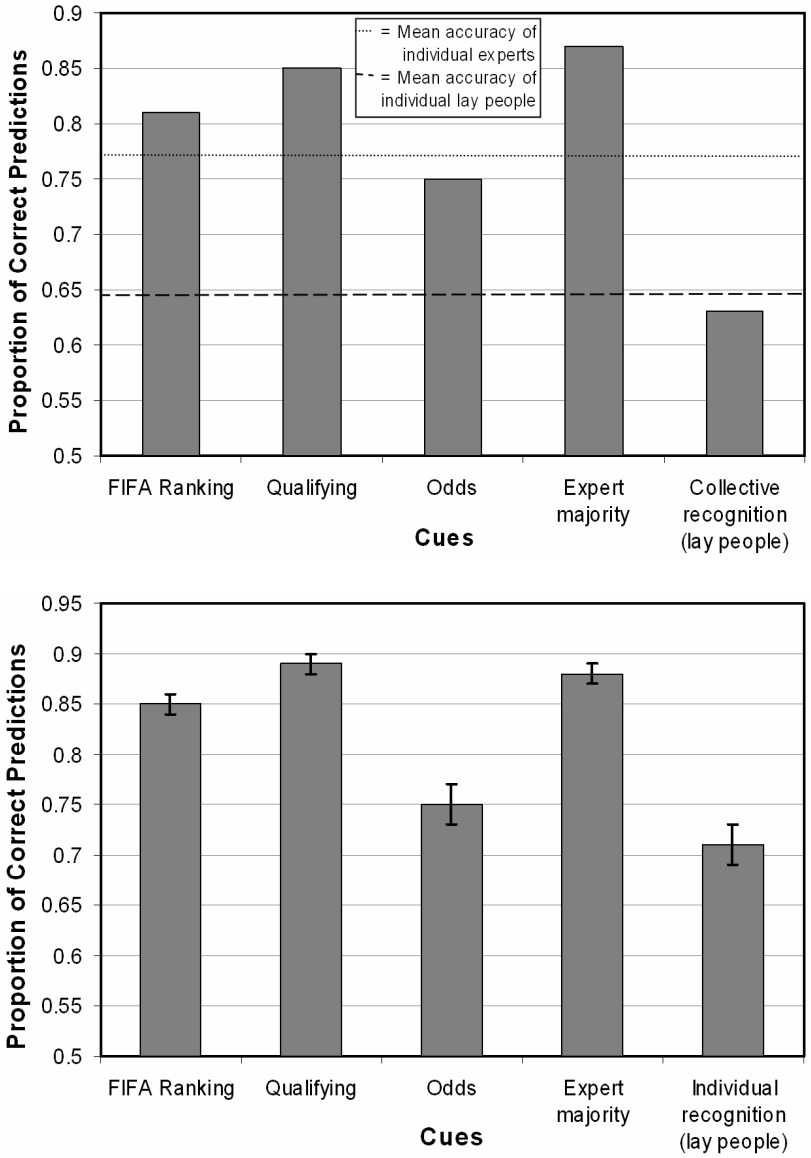

Figure 2.2.4. (a) Proportion of correct predictions of the different direct indicators of team strength of the actual outcome of all 24 games compared to the proportion of correct predictions of (collective) recognition. The mean accuracies of the individual expert and lay participants are indicated by the two dashed lines. (b) Mean proportion of correct predictions (across participants) of the different indicators when one team was recognized but not the other. The bars indicate standard errors.

Discussion

In this chapter, I tested the recognition heuristic as a model of how people with limited knowledge make forecasts of sports events. Furthermore, I investigated the less-is-more effect, a possible consequence of using the recognition heuristic. Three questions were addressed. First, how well does the recognition heuristic describe laypeople's forecasts for matches at the EURO 2004 compared to other plausible mechanisms? The recognition clearly outperformed the other mechanisms. Second, is there a less-is-more effect across people with

different levels of knowledge? No less-is-more effect was found between the lay and expert groups (because experts had a higher knowledge validity than the laypeople), but a less-is-more effect emerged across lay participants. This less-is-more effect, however, was shifted, apparently due to the number of teams recognized (*n*) being correlated both with recognition validity α and knowledge validity β. Finally, how well does recognition predict success compared to more direct indicators of team strength, and how well do experts predict relative to simple statistical models that are based on these indicators? Concerning the recognition heuristic, this comparison showed that although recognition was able to predict well above chance level, it lagged behind the direct indicators. As to the accuracy of the experts, simple statistical models that use only one indicator outperformed the average accuracy of the individual experts. In the following, I discuss the implications of these findings.

People's Use of the Recognition Heuristic

Study 3 corroborates previous findings (Ayton & Önkal, 2004; Serwe & Frings, 2004) that the recognition heuristic represents an important tool people with limited knowledge use when making forecasts of sports events. I show that when compared to alternative models of lay prediction, the recognition heuristic does best in accounting for people's forecasts. However, it could be argued that the alternative lay mechanisms were at a disadvantage as it was merely assumed, but not certified, that laypeople knew the relevant information (GDP, population size, etc.). Although it is possible that some participants did not know the correct rank order of the countries on the relevant dimensions, by using a relatively high difference threshold (20%), only weak assumptions were made concerning the precision of people's general knowledge. In addition, the recognition heuristic still described the lay predictions best when collective recognition, rather than individual recognition, was used. Of course, one cannot exclude that mechanisms other than those considered would have fared better, but, apart from the fact that a proportion of 90% correctly predicted forecasts is difficult to exceed, the four models that were tested are plausible ones and indeed useful to predict the actual outcomes (see Table 2.2.3).

What might have driven people's selection of the recognition heuristic? As Table 3 shows, recognition had the highest validity of the candidate mechanisms, closely followed by EAST–WEST. However, it also had the lowest discrimination rate. This suggests that the high validity of recognition was the crucial determinant of strategy selection. Thus, in contrast to work by Newell, Rakow, Weston, and Shanks (2004) suggesting that people consider cues according to their "success" rate (which represents a combination of validity and DR), it appears that the discrimination rate was of little importance in the strategy selection here.

However, as mentioned above, the validity of EAST–WEST was inferior to recognition only by a small margin. Moreover, both mechanisms represent, in the context studied, inference from memory (Gigerenzer & Todd, 1999) and are based on information arguably available to the participants. Therefore, one might ask why EAST–WEST was not

used more. Though speculative, one plausible explanation is that recognition is "cheap". The judgment of whether one has heard of the name of an object before is generated more or less automatically when an object name is processed (e.g., Grill-Spector & Kanwisher, 2005; Thorpe, Fize, & Marlot, 1996), whereas more effort is required to retrieve the information considered by the other mechanisms (see Chapter 2.1).

Regarding the applicability of the recognition heuristic, as mentioned above, this heuristic could not be used throughout. This is because the heuristic relies on very "coarse" information (with dichotomous cues such as recognition representing the extreme high end of coarseness). In a reference class of N objects, the recognition cue can have a maximum DR of 1 only with $N = 2$ objects and with increasing N, the maximum DR quickly converges to 0.5 (Gigerenzer & Goldstein, 1996). This means that in many situations the recognition heuristic can at most be applied in only around half of all possible comparisons. It was not in the scope of this study to investigate what mechanisms are used when the recognition heuristic cannot be applied (e.g., when both teams are recognized). One possibility could be lexicographic strategies such as Take The Best (Gigerenzer & Goldstein, 1996), which looks up cues in decreasing order of validity, or the fluency heuristic (Schooler & Hertwig, 2005), which uses recognition speed to discriminate between two recognized objects. Future research should pursue such models as possible strategies for making forecasts of sports events.

The Ecological Rationality of Using Recognition for Predicting Sports Events

Recognition allowed participants to predict the match outcomes well above chance level (71% and 63% correct forecasts for individual and collective recognition, respectively), but it was considerably worse than all direct indicators of team strength that were tested. This result is at odds with the findings of Serwe and Frings (2004), who found, in the context of tennis matches, that recognition performed similar to rankings and odds. It should be noted, however, that in the present study recognition performed similarly as in Serwe and Frings' study in absolute terms (with 71% vs. 73-76% correct predictions), whereas the performance of rankings I found was much higher (85 and 89% vs. 68%; see Figure 2.2.4b). The inferiority of recognition in the present study is thus mainly due to the rankings reaching higher accuracy than in the tennis domain studied by Serwe and Frings. One speculative reason for this might be that the rankings of soccer teams are more robust than ATP rankings of the tennis players.

Less Is More: What the Correlation between n and α Tells Us about How Objects Are Learned

One noteworthy feature of the recognition heuristic is that it provides a straightforward explanation for the less-is-more effect, which was observed here and in previous studies on forecasting sports events (Andersson et al., 2003, 2005; Ayton & Önkal, 2004). In their original description of the recognition heuristic, Goldstein and Gigerenzer (2002) described the conditions under which the less-is-more effect is expected to occur: the recognition

validity α must be larger than the knowledge validity β. In the present study it was found that the less-is-more effect can be shifted (i.e., emerge only at very high levels of knowledge), and this points to a necessary specification of this condition for the effect to emerge. If the number of recognized objects is correlated with α and/or β, it is possible that there will be no less-is-more effect even when $\alpha > \beta$. Note that in the present study the effect was markedly shifted although the correlations between *n* and α and β were quite small ($r = .19$ and .22, respectively). With higher correlations the effect could disappear completely.

But why is *n* correlated with α and β? That β is correlated with *n* is not too surprising, as learning about more teams often comes with accumulating further knowledge about already known teams, which helps in making better forecasts. More interesting is the correlation between α and *n*. This correlation means that with higher *n*, the relationship between recognition and the strength of a team becomes stronger. One way to interpret this correlation is that it might hint at how the teams are sequentially learned. More specifically, the correlation suggests that among the teams with a very high probability of being recognized (i.e., those that are likely to be recognized even with low *n*) there is a substantial proportion of teams of only intermediate strength. As a consequence, α is only modest with low *n*. Conversely, the teams with a low probability of being recognized (which are thus likely to be recognized only at high levels of *n*) are almost exclusively teams of low strength, resulting in a higher α when *n* increases. Thus, with very little knowledge one should not be too confident that the recognized teams are also successful, whereas with a lot of knowledge, the lack of knowledge of a team is very informative of the strength of a team. Put differently, the imperfect correlation between recognition and strength of a team is mainly due to unsuccessful teams being recognized rather than to successful ones being unrecognized.

Study 3 thus demonstrates the influence of the correlation between *n* and α and β on the emergence of a less-is-more effect. Specifically, the condition $\alpha > \beta$ is insufficient for the effect to occur. Moreover, if α and β vary systematically among people, the condition $\alpha > \beta$ is not only insufficient, but even unnecessary. If α and/or β are *negatively* correlated with *n* (i.e., α and/or β are *decreasing* with increasing *n*), there can be a less-is-more effect even when, on average, $\alpha < \beta$. Whether and when such a negative correlation occurs in the real world is currently unknown, but the results obtained here clearly point to the importance of examining this correlation.

Accuracy of Expert versus Lay Prediction

Whereas in previous studies predictions by sports experts were often no more accurate than those by non-experts (e.g., Andersson et al., 2003; 2005), Study 3 showed clearly superior forecasts by the experts. Why? As pointed out by various authors, an analysis of the task environment—that is, how predictive the available cues are—is important if one aims to understand expert performance (Shanteau, 1992; Stewart, Roebber, & Bosart, 1997). Often, the cues available to experts will be more valid than cues available to laypeople. The analysis

of the (admittedly assumed) task environment of experts (i.e., the direct indicators of team strength) suggests that this was indeed the case in the present study: the direct indicators of team strength were more predictive of the actual results than recognition, the information hypothesized to be used by laypeople. As a consequence, the experts could achieve a much higher accuracy than in comparable studies (e.g., more than ten percentage points better than in Ayton and Önkal, 2004).

One reason for the often reported failure of experts to clearly outperform laypeople might then be that there are situations in which the cues used by laypeople are more predictive than the cues used by experts. Moreover, it is even possible that laypeople equal experts if laypeople have fewer valid cues available. This can happen due to the phenomenon known as overfitting. Overfitting occurs when experts integrate multiple cues[35] and choose weights for this integration that are optimal in a given situation but fail to generalize to new situations. It is likely that cue weights do not generalize when they are tuned to explain outcomes that are in fact due to random error, thus fitting noise. In terms of the forecasting task, an expert might post hoc be able to explain all results of a tournament by creating a model that considers a multitude of cues, but this model would fare rather badly when used to predict the results of a future tournament. Accordingly, Czerlinski, Gigerenzer, and Goldstein (1999) showed that simple rules using only little information are more robust than information-greedy strategies such as multiple regression (see also Dawes, 1979). In the same vein, overfitting might also explain why experts, as in Study 3, often perform worse than "naïve" strategies that rely on only one cue (Boulier & Stekler, 2003; Forrest & Simmons, 2000). Therefore, in situations in which experts cannot exploit available information—because they use the wrong weights—simple strategies such as the recognition heuristic may prove more robust, with the result that experts and laypeople end up being equally accurate, while using different strategies.

Conclusion

Limited knowledge does not necessarily imply bad forecasts. In the sports context studied in this chapter, people with only very limited soccer knowledge used the fact that they had not heard of some teams as an indication that these teams were not as strong as those they had heard of. Although this recognition knowledge could not predict the actual results as well as various direct indicators of team strength, it allowed participants to make correct predictions well above chance level and fared better than all the other plausible lay mechanisms that were tested. As a consequence of the superior validity of recognition relative to the validity of further general knowledge that has to be recruited when recognition does not discriminate, I observed that people who recognized fewer teams made more correct predictions than people who recognized (almost) all teams. When ignorance is systematic, incomplete knowledge pays.

[35] Indeed, existing research on predictions by sports experts suggests that experts consider a multitude of cues (Boulier & Stekler, 2003; Forrest & Simmons, 2000).