CHAPTER 3
Recall: The Social Environment as the Window to the World

3.1 Judgments of Risk Frequencies: Tests of Possible Cognitive Mechanisms

How the public perceives health risks has been a long-standing concern in the medical community. In light of the potential risk of old and ever new emerging epidemics (such as SARS or bird flu) it is more important than ever to shed light on the psychological mechanisms underlying the public's perception of health risks. This is the goal of this chapter. Specifically, the focus is on one dimension of risk perception, namely, people's assessment of risk frequencies. Previous research in psychology has offered two rather contradictory views of how, and how accurately, people estimate the frequencies of events. These views were pitted against each other to assess their relative merits. After formulating two cognitive mechanisms implied by each view and deriving specific predictions from each mechanism, I tested the predictions at the level of aggregate frequency judgments and estimates (Studies 4 and 6) and at the level of individual frequency judgments (Study 5).

Judging Risk Frequencies: Heuristic Inference or Direct Encoding?

Calibrating oneself to all the risks in one's environment is a task of Herculean proportions. For instance, there are currently more than 1,400 documented microorganisms that can infect humans, and this is just the tip of the iceberg: Only an estimated 1% of the bacteria and 4% of the viruses on the planet have been identified thus far (Glasser, 2004, p. 36). Infections, in turn, represent only one class of health risks. Among the many others are the risks posed by artifacts such as guns, cars, and electric outlets; by natural hazards such as tornadoes, floods, and lightning; and by human carcinogens such as asbestos, solar radiation, and tobacco smoking.

How do real people—that is, people constrained by limited time, limited memory, and limited computational capacities—judge the frequency of risks in their environment, and how well do they do it? Research in psychology on how people estimate the frequency of events has given rise to two very different views on these questions. One view suggests that event frequencies are tracked directly and that the automaticity of the tracking process allows for impressively accurate frequency estimates. At least implicitly rejecting the premise that frequency estimates are based on directly retrievable frequency records, the other view holds that people infer the distal criterion (i.e., event frequency) by exploiting a proximal cue, namely, *availability*. Although often appropriate, reliance on this cue to judge frequency can lead to systematic biases in risk perception. Next, I describe both of these accounts of frequency judgments in detail, beginning with the notion of availability.

Availability Heuristic

Tversky and Kahneman (1974), who first proposed the *availability heuristic,* characterized it thus:

> There are situations in which people assess the frequency of a class or the probability of an event by the ease with which instances or occurrences can be brought to mind. For example, one may assess the risk of heart attack among middle-aged people by recalling such occurrences among one's acquaintances. (Tversky & Kahneman, 1974, p. 1127)

Availability was the key explanatory concept in a seminal study by Lichtenstein, Slovic, Fischhoff, Layman, and Combs (1978) on judgments of risk frequency. They asked participants to judge the mortality rate (in the United States) associated with a wide range of risks, including motor vehicle accidents, poisoning by vitamins, and lung cancer. Frequency judgments were elicited from each participant in two ways: Presented with a pair of risks, participants were first asked to say of which risk a randomly selected person would be more likely to die and to estimate how many times more likely a person would be to die of this risk as opposed to the other risk. Other participants were required to estimate the mortality rate attributable to each individual cause of death in an average year.

In reviewing their own and related studies, Slovic, Fischhoff, and Lichtenstein (1982) emphasized that "because frequently occurring events are generally easier to imagine and recall than are rare events, availability is often an appropriate cue" (p. 465) to event frequency. Availability is not a foolproof cue, however, because it is also affected by factors that are unrelated or even negatively related to event frequency, such as "disproportionate exposure, memorability, or imaginability" (Lichtenstein et al., 1978, p. 551). For instance, a moviegoer who has just watched *Jaws* (Zanuck, Brown, & Spielberg, 1975) would likely have little trouble imagining the occurrence of a shark attack and might therefore overestimate its probability, which is objectively low.[36]

As a result of such potential dissociations between frequency of occurrence and availability in memory, risk frequency judgments can be systematically distorted. Specifically, Lichtenstein et al. (1978) identified two major biases that they attributed to the availability heuristic.

The *primary bias* is the "overestimation of low frequencies and underestimation of . . . high frequencies" (Lichtenstein et al., 1978, p. 574) in people's estimates of mortality rates. Figure 1 illustrates this effect by plotting participants' average frequency estimates against the actual frequencies from public health statistics. Whereas the average estimated frequencies of relatively rare events (such as botulism and tornadoes) are larger than the actual frequencies, the average estimated frequencies of common events (such as stroke and diabetes) are smaller than the actual frequencies. The *secondary bias* refers to the observation that "different pairs

[36] According to the Florida Museum of Natural History's shark research Web site (http://www.flmnh.ufl.edu/fish/Sharks/sharks.htm), four fatalities occurred in 2003 worldwide.

[of causes of death] with the same [probability] ratio had quite different judged ratios" (Lichtenstein et al., 1978, p. 558). For instance, deaths due to motor vehicle accidents are only 1.5 times more frequent than deaths caused by diabetes; Lichtenstein et al.'s (1978) college students, however, estimated the former to be an average of about 350 times more frequent than the latter.

How can the availability heuristic explain the primary and secondary biases? According to Lichtenstein et al. (1978), the primary bias arises when two conditions hold: (a) People base their estimates on recalled instances, and (b) the number of recalled instances is largely independent of the actual frequency of the event—an assumption for which Lichtenstein et al. marshaled support by referring to B. H. Cohen (1966). Consequently, it is possible that people recall as many cases of death from measles as of death from diabetes among their acquaintances despite the fact that the latter event is much more frequent than the former. Lichtenstein et al. explained the secondary bias by proposing that the ease with which instances of an event can be brought to mind or recalled is affected by the event's vividness. Whereas some risks represent "undramatic, quiet killers," others represent "sensational events" (Lichtenstein et al., 1978, p. 575), and the latter can by more easily brought to mind.

Lichtenstein et al.'s (1978) explanation of risk frequency judgments in terms of the availability heuristic has been more or less taken for granted since it was proposed (e.g., Folkes, 1988; MacLeod & Campbell, 1992; Stapel, Reicher, & Spears, 1994; Sunstein, 2002). Yet neither Lichtenstein et al. nor later researchers tested specific predictions derived from the heuristic. Instead, the heuristic was typically invoked as a post hoc explanation for the findings. In addition, the actual mechanism of availability was left ambiguous in Tversky and Kahneman's (1973) original paper. As has frequently been pointed out (e.g., Betsch & Pohl, 2002; Brown, 1995; Fiedler, 1983; Schwarz & Wänke, 2002), Tversky and Kahneman's formulation of availability is consistent with two different mechanisms—one that is based on the amount of actually recalled instances and one that is based on the (anticipated or experienced) ease of recall. I propose the following definitions of these mechanisms.

Availability-by-recall mechanism. In the context of risk frequency judgments, I define availability by recall as the number of deaths due to specific risks that one recalls having occurred in one's social circle, by which I mean one's family, friends, and acquaintances. Using availability by recall, one judges whether more people die of heart attacks or breast cancer, for example, by retrieving from memory specific cases of death from heart attack and breast cancer, respectively, within one's social circle. The number of recalled instances serves as a cue to the criterion (i.e., the mortality rate associated with each risk in the population).[37]

[37] Benjamin and Dougan (1997) have argued that in the context of health and safety risks, consideration of risk events in one's social environment represents an adaptive strategy when assessing risks. Furthermore, they showed that such a sensitivity to occurrences among one's age cohort is reflected in Lichtenstein et al.'s (1978) original data.

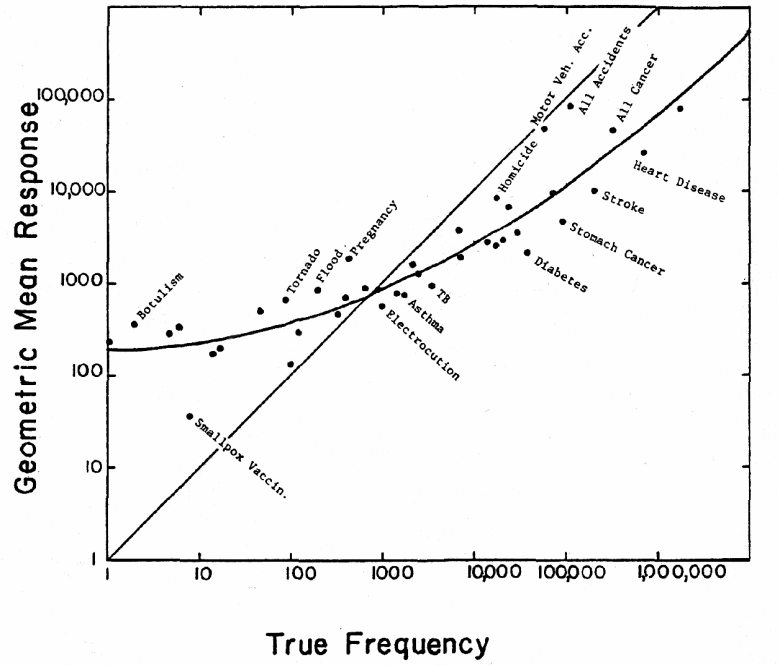

Figure 3.1.1. The primary bias, illustrated by the relationship between estimated and actual number of deaths per year for 41 causes of death in Lichtenstein, Slovic, Fischhoff, Layman, and Combs (1978). Each point is the mean estimate (geometric mean) of 39 students. The observation that, for rare causes of deaths, the mean estimated number is higher and that, for frequent causes, this number is lower has been called the primary bias. The curved line is the best fitting quadratic regression line.

Fluency mechanism. This mechanism is inspired by the assumption that in judging availability, "it is not necessary to perform the actual operations of retrieval" (Tversky & Kahneman, 1973, p. 208); it suffices to anticipate the ease with which relevant instances could be brought to mind. For instance, one judges whether more people die of heart attacks or breast cancer by assessing the ease with which such instances could be brought to mind without actually retrieving them. This subjective judgment of ease of retrieval serves as a cue on whose basis the frequency of each risk can be inferred. Although ease of retrieval has been effectively manipulated in recent studies (e.g., Schwarz & Vaughn, 2002), it has rarely been measured (but see Sedlmeier, Hertwig, & Gigerenzer, 1998).

One way to define ease of retrieval is by relating it to the notion of fluency of processing of an object once it has been encountered (see, e.g., Jacoby & Brooks, 1984; Toth & Daniels, 2002; Whittlesea, 1993). In fact, Jacoby, Kelley, Brown, and Jasechko (1989) explicitly articulated the link between availability and fluency:

> Reading a word once allows it to be read more fluently later. . . . An item seems familiar if it can be easily brought to mind or fluently processed. This account of familiarity in terms of fluency is analogous to Tversky and Kahneman's account of probability estimations based on an availability heuristic. (Jacoby et al., 1989, p. 328)

In numerous studies, processing fluency—mediated by prior experience with a stimulus—has been shown to function as a cue in a range of judgments. For example, more fluent processing due to previous exposure can increase the perceived fame of nonfamous names (the false fame effect; Jacoby et al., 1989) and the perceived truth of repeated assertions (the reiteration effect; Begg, Anas, & Farinacci, 1992; Hertwig, Gigerenzer, & Hoffrage, 1997).

In my second interpretation of availability (henceforth referred to as the *fluency mechanism*), I assume that previous experience with a stimulus, such as a word denoting a risk, increases the fluency with which the stimulus is later processed and that fluency of processing is associated with the ease with which occurrences of the respective risk can be retrieved. I therefore define ease of retrieval in terms of the frequency with which words such as *heart attack, homicide,* and *botulism* have been encountered. Of course, this raises the question of how to determine the frequency of encounters with words. In my view, one elegant proxy is environmental statistics—that is, using tallies of the frequencies with which such words appear in print media as a proxy for the frequency of encounters with words.

Direct Encoding

Viewed in light of an influential research program launched by Hasher and Zacks (1979, 1984), calibrating oneself to risk frequencies in one's environment may not be the Herculean task that it initially appears to be. On the basis of their studies demonstrating people's "pervasive sensitivity" to event frequencies, these authors proposed that frequency information enters memory via an encoding mechanism that automatically processes "fundamental attributes of experience" such as spatial location, temporal order, and frequency of occurrence (Zacks & Hasher, 2002, pp. 22, 25). In this framework, automatic encoding means that the encoding of, for instance, frequency information makes minimal demands on attentional resources and does not require intention.

Hasher and Zacks's (1984) automatic encoding thesis has been extensively tested (for reviews, see Barsalou, 1992, and Zacks & Hasher, 2002). In response to these tests, Zacks and Hasher (2002) proposed the following modification of the automaticity claim, which gives attention a key role: "The encoding of frequency information is an inevitable consequence of attending to events, and in that sense, is obligatory" (p. 34). Regardless of its processes, however, information encoding appears to result in highly accurate frequency estimates. As Jonides and Jones (1992) put it, "Ask about the relative numbers of many kinds of events, and you are likely to get answers that reflect the actual relative frequencies of the events with great fidelity" (p. 368; see also Zacks & Hasher, 2002, p. 27). Using their conclusion as a starting point, I now present two mechanisms of how people could make risk judgments on the basis of directly encoded frequency information.

Regressed-frequency mechanism. In the context of risk frequency judgments, the regressed-frequency mechanism assumes (a) that people monitor the occurrence of individual health risks (e.g., based on personal experiences, the reading of obituaries, media reports,

physicians' warnings, and public awareness campaigns) and (b) that in light of unreliability of processing, the estimated mortality rates are regressed toward the mean such that small frequencies are overestimated and large frequencies are underestimated (Fiedler & Armbruster, 1994). In contrast to Lichtenstein et al.'s (1978) view, which assumes that people have biased knowledge of risk frequencies because of "disproportionate exposure, memorability, or imaginability of various events" (p. 551), the regressed-frequency mechanism assumes that people's frequency knowledge is roughly accurate except for the estimates' tendency to regress toward the mean. It should be noted that not only is this tendency akin to the primary bias observed by Lichtenstein et al. but also it is ubiquitous in studies of other types of frequency judgments (e.g., Begg, Maxwell, Mitterer, & Harris, 1986; Greene, 1984; Hintzman, 1969, 1988; Sedlmeier et al., 1998; Shanks, 1995; Williams & Durso, 1986; Zacks & Hasher, 2002).

Risk-category mechanism. When one learns that a neighbor has passed away, one may not learn the exact cause of his or her death. For instance, one may be told that the neighbor died of cancer but never find out the precise type of cancer from which he or she suffered. The event is thus inexactly represented. On the basis of the premise that such inexact representations are the rule rather than the exception, the risk-category mechanism postulates that the frequency of specific events is judged by reference to the central value of the category to which they belong. For example, a person who does not know the mortality rates associated with lightning and ovarian cancer may nevertheless have the (accurate) sense that the average mortality rate for the category natural hazards is markedly lower than the average mortality rate for the category diseases. Therefore, the person judges death from ovarian cancer to be more likely than death from lightning.

Several authors have espoused the thesis that information about the superordinate categories of an object is used to judge individual objects (e.g., Brown, 2002b; Fiske & Pavelchak, 1986). For instance, in Huttenlocher, Hedges, and Vevea's (2000) category adjustment model, estimates of the value of a stimulus on a dimension are a blend of both fine-grained information about the stimulus and knowledge derived from the category (e.g., the shape of the distribution or the central tendency of values) to which the stimulus belongs. The higher the uncertainty regarding the fine-grained information, that is, the less exact the stimulus representation, the more weight the category information is given when deriving an estimate. In the extreme case of complete lack of stimulus-specific information, the estimate for the stimulus coincides with the central tendency of the category.

It is interesting to note that Huttenlocher et al.'s (2000) model predicts overestimation of small stimulus values and underestimation of large stimulus values—the very phenomenon Lichtenstein et al. (1978) referred to as primary bias. However, in the category adjustment model, this phenomenon is seen as the side effect of a normative judgment strategy that aims to minimize error in light of uncertain knowledge.

Predictions

In what follows, I derive specific predictions for each of the four mechanisms (i.e., availability by recall, fluency, regressed frequency, and risk category). The predictions assume a context in which people are given two risks and asked to decide which one is more frequent. To explore how robust the mechanisms' performance would be across different health risk environments, I tested their predictions using different sets of risks and different target criteria (i.e., mortality rate and disease incidence). The first set encompassed the causes of death that Lichtenstein et al. (1978) examined. Table 3.1.1 lists them and their respective mortality rates in Germany. I refer to this set as the *assorted set* because it compiles risks across various categories. Two other sets included all types of cancer and all notifiable infectious diseases in Germany, respectively. I refer to these sets as the *cancer set* and the *infection set*. Table 3.1.1 lists the events in both sets and the respective incidence rates in Germany. For the latter two sets, participants' target criterion was the diseases' annual incidence rates. It is worth mentioning that the cancer set and the infection set rested on existing classifications, that is, the decision of which events to include was not mine; in contrast, in the assorted set, I adopted the composition chosen by Lichtenstein et al. (1978, p. 554).[38]

In addition, the cancer set and the infection set, unlike the assorted set, did not include entries with different degrees of abstraction (e.g., all disease) that may have invited different inductive or deductive strategies.

Availability by Recall

This mechanism assumes that the choice between two risks is a function of the actual recall of deaths (or instances of diseases) among one's social circle. To be able to specify the predictions for specific risks, a pilot study was conducted to obtain numerical values. Forty participants were presented with the events in the assorted set. For each cause of death, they were asked to recall occurrences of deaths in their social circle (i.e., family, friends, and acquaintances) and to write down the number of instances they could retrieve. Similarly, two groups of 60 participants each were presented with the infection set and the cancer set and asked to recall occurrences of instances of such diseases in their social circle. This recall task rendered it possible to specify predictions of the availability-by-recall mechanism for individual pair comparisons. The recall data also provided a test for Lichtenstein et al.'s (1978) assumption that actual recall is largely independent of the frequency of the event (see above). Contrary to this assumption, the number of recalled cases for each risk in the assorted set was strongly correlated with the actual frequencies (Spearman rank correlation = .77). In the cancer set and the infection set, the correlations amounted to .61 and .43, respectively.

[38] Previous follow-ups of the Lichtenstein et al. (1978) studies have mostly focused on the assorted set (e.g., Benjamin, Dougan, & Buschena, 2001; Carnegie Mellon University Graduate Research Methods Class, 1983).

Table 3.1.1. The entries in the assorted set, the cancer set, and the infection set, the entries' average annual mortality rates (assorted set) and incidence rates (cancer and infection sets), respectively (averaged rates for the years 1996–2000), and the median estimated frequencies (Study 6).

	Annual rate	Median estimate		Annual rate	Median estimate
Assorted set			*Cancer set*		
Fireworks	0	30	Penis cancer	551	1,000
Flood	0	30	Bone cancer	939	5,200
Whooping cough	0	10	Cancer of the connective tissue[a]	1,216	2,500
Smallpox	0	9	Thyroid cancer	2,987	5,000
Smallpox vaccination	0	5	Larynx cancer	3,084	5,500
Tornado	0	0	Testicular cancer[a]	3,439	6,000
Poisoning by vitamins	0	50	Esophageal cancer[a]	3,821	4,000
Measles	2	15	Hepatic cancer[a]	4,835	5,000
Polio	3	40	Cancer of the gall bladder	5,489	3,000
Lightning	7	10	Skin cancer	6,563	25,000
Firearm accident	19	100	Cancer of the nervous system[a]	6,931	11,500
Venomous bite or sting	20	80	Ovarian cancer[a]	7,819	6,000
Syphilis	24	11	Cancer of the mouth and throat	10,273	3,900
Nonvenomous animal	26	30	Pancreatic cancer	10,315	5,000
Pregnancy, childbirth, and abortion	45	150	Renal cancer	13,036	3,000
Motor vehicle-train collision	48	100	Bladder cancer[a]	15,368	2,500
Botulism	74	100	Cervical cancer	16,478	13,450
Electrocution	93	200	Stomach cancer	18,252	9,000
Excess cold	159	39	Rectal cancer	20,981	4,000
Appendicitis	242	100	Leukemia and lymphoma[a]	23,937	15,000
Infectious hepatitis	321	250	Prostate cancer	29,681	12,000
Poisoning by solid or liquid	493	500	Colon cancer	33,373	8,000
Fire and flames	526	200	Lung cancer[a]	36,964	36,000
Drowning	538	51	Breast cancer[a]	46,248	35,000
Tuberculosis	551	100			
Homicide	800	1,000	*Infection set*		
Emphysema	2,790	398	Poliomyelitis	0.25	300
Asthma	4,086	250	Diphtheria	1	1,000
Leukemia	6,844	1,500	Egyptian ophthalmia/trachoma	1.75	691
Accidental falls	7,985	1,000	Tularemia/rabbit fever	2	200
Motor vehicle (car, truck, or bus) accidents	8,028	13,500	Cholera	3	17.5
			Leprosy[a]	5	0.75
Suicide	11,670	1,603	Tetanus	9	1,000
Breast cancer	18,249	4,000	Hemorrhagic fever[a]	10	150
All accidents	20,784	80,000	Botulism/food poisoning[a, b]	15	37,500
Diabetes	21,820	400	Trichinosis	22	326.5
Lung cancer	37,728	8,000	Brucellosis/undulant fever	23	146.5
Stroke	47,276	10,000	Leptospirosis/Well's disease[a]	39	370
Cancer of the digestive system	69,744	8,000	Gas gangrene	98	400
All cancer	211,467	107,693	Ornithosis/parrot fever	119	225
Heart disease	410,869	50,000	Typhoid and paratyphoid[a]	152	200
All disease	783,645	350,000	Q fever	179	200
			Malaria	936	400
			Syphilis[a]	1,514	1,500
			Bacterial dysentery/Shigellosis	1,627	1,000
			Gonorrhoea[a]	2,926	6,000
			Meningitis and encephalitis	4,019	5,000
			Tuberculosis[a]	12,619	1,500
			Viral hepatitis[a]	14,889	10,000
			Gastroenteritis (infective enteritis)[a]	203,864	37,000

[a] Included in Study 4.

[b] Not included in analysis (see footnote 43).

Note: The mortality rates for the assorted set were taken from tables made available by the Federal Statistical Office of Germany for the years 1996 to 2000 (e.g., Statistisches Bundesamt, 2002). The incidence rates for the cancer and infection sets were taken from tables made available by the Robert Koch Institute for the years 1997 to 2000 (Arbeitsgemeinschaft Bevölkerungsbezogener Krebsregister in Deutschland, 1999). The infection set encompassed 24 infections ("dangerous infectious diseases"; see e.g., Robert Koch Institute, 2001) that by law are notifiable in Germany ("Bundesseuchengesetz"—a law that has recently been revised and now encompasses, for instance, HIV). Note that "rabies" was dropped from the infection set because there was no single incident during the specified time period.

Availability by recall assumes that the choice between two risks is a function of the number of cases (deaths or cases of disease) recalled from participants' social circles (as defined above), and its prediction can be stated as follows:

$$\text{Choice proportion}_{\text{Risk } a} = \sum\text{Recalled instances}_{\text{Risk } a}/(\sum\text{Recalled instances}_{\text{Risk } a} + \sum\text{Recalled instances}_{\text{Risk } b}),$$

where $\text{Choice proportion}_{\text{Risk } a}$ is the proportion of participants who select Risk *a* to be more likely than Risk *b,* and $\sum\text{Recalled instances}_{\text{Risk } a}$ and $\sum\text{Recalled instances}_{\text{Risk } b}$ are the sum of instances (recalled by all participants) of Risk *a* and Risk *b,* respectively. Here and throughout this chapter, Risk *a* denotes the event that is, in reality, the more frequent one in a pair comparison. It was not simply assumed that if, on average, more instances of Risk *a* than Risk *b* were recalled, then 100% of participants would choose Risk *a*. Rather than using such a deterministic rule, I employed a probabilistic choice rule to derive choice proportions. That is, it was assumed that the probability that *a* would be chosen was proportional to *a*'s relative support (i.e., the ratio of the recalled instances for Risk *a* over the sum of the recalled instances for Risks *a* and *b*).

Fluency

The fluency mechanism assumes that the choice between two risks is a function of the fluency with which the names of the risks are processed when they are encountered. As a proxy for ease of retrieval and fluency, I determined how often the terms denoting causes of death and diseases were mentioned in German print media. Using COSMAS I, an extensive data archive of German daily and weekly newspaper articles, I counted the frequency of occurrence with which, for instance, the words *died from breast cancer* were mentioned.[39] They occurred 3,302 times. I did the same for all causes of death in the assorted set and for all events in the cancer and infection sets. For the latter sets, I used only the names of the diseases (excluding *died from*). I found that the rank correlations between the number of mentions of a risk and its actual frequency were .74, .44, and .23 in the assorted, the cancer, and the infection sets, respectively.

The fluency mechanism assumes that the choice between two risks is a function of their number of mentions. Its prediction can thus be stated as follows:

$$\text{Choice proportion}_{\text{Risk } a} = \sum\text{Occurrences}_{\text{Risk } a}/(\sum\text{Occurrences}_{\text{Risk } a} + \sum\text{Occurrences}_{\text{Risk } b}),$$

where $\sum\text{Occurrences}_{\text{Risk } a}$ and $\sum\text{Occurrences}_{\text{Risk } b}$ are the number of mentions of Risk *a* and Risk *b,* respectively

[39] 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 1,211,000,000 words.

Regressed Frequency

According to this mechanism, people keep track of the frequency of occurrences of individual health risks. Thus, their frequency judgments conform to the actual frequencies of events except that the estimates tend to regress toward the mean frequency within the set of risks (i.e., low frequencies are overestimated, and high frequencies are underestimated). The amount of regression was assumed to be 10%. To arrive at this estimate, I analyzed the risk frequency judgments observed by Christensen-Szalanski, Beck, Christensen-Szalanski, and Koepsell (1983). They asked experts (physicians) and nonexperts (students) to estimate mortality rates of various diseases. The results from the latter group were used to estimate the amount of regression because the focus here was on lay judgments. The median amount of regression observed in students' estimates was 10.2%.[40]

On the basis of this amount of regression, the prediction of the regressed-frequency mechanism can be stated as follows:

$$\text{Choice proportion}_{\text{Risk } a} = \text{Regressed frequency}_{\text{Risk } a}/(\text{Regressed frequency}_{\text{Risk } a} + \text{Regressed frequency}_{\text{Risk } b}),$$

where the regressed frequencies are the actual mortality rates or incidence rates of Risk *a* and Risk *b,* respectively, regressed by the factor 0.1.[41]

Risk Category

The risk-category mechanism assumes that the frequency estimate for an individual risk is inferred from the average frequency in the category to which the risk belongs. Lichtenstein et al.'s (1978) original list included at least three such categories of risks, namely, diseases, accidents, and natural hazards.[42] In Germany, the average mortality rates in these three categories were 4,835, 860, and 25, respectively. That is, many more people died on average from diseases than from accidents, and more people died from accidents than from natural hazards. In addition, the assorted set included not only individual risks (e.g., breast

[40] To determine the amount of regression, I followed the procedure used by Sedlmeier et al. (1998). First, both the actual frequencies of the diseases and the geometric mean judgments were transformed to percentages (of the 42 diseases in Christensen-Szalanski et al., 1983, I excluded 7 as no definite actual frequencies were reported). That is, the absolute values were expressed in relation to the sum of all frequencies (sum of actual frequencies for all diseases = 100%; sum of mean judgments for all diseases = 100%). As a result of this transformation, both actual and judged frequencies had an identical mean (100% divided by 35 diseases = 2.86%). Next, the distances of both the transformed actual frequencies and the transformed mean judgments from this mean were calculated, yielding the distance measures *AD* and *JD* for the actual frequencies and the mean judgments, respectively. Finally, the amount of regression of the judgments for each disease was determined by $100 - (JD/AD) \times 100$. This value is zero if the deviation from the mean of the judged frequency equals the actual frequency ($JD/AD = 1$). It is positive if the deviation is smaller, that is, if there is a regression effect ($JD/AD < 1$), and it is negative if the deviation is larger ($JD/AD > 1$). Across all events, the median amount of regression was determined.

[41] The value of, say, breast cancer was calculated as follows: $\text{Regressed actual frequency}_{\text{breast cancer}}$ = actual mortality rate in the assorted set – 0.1 × (actual mortality of breast cancer – average mortality rate in the assorted set).

[42] Note that each category subsumes multiple subcategories: The category of accidents, for instance, includes 24 subcategories, according to *ICD-10* (World Health Organization, 1992), using the two-digit codes.

cancer or firearm accidents) but also summation categories such as all disease, all cancer, all accidents, suicide, and homicide. For these summation categories, it was assumed that the frequency judgments were a function of the total sum in the respective categories. Specifically, for the total of eight categories (diseases, accidents, natural hazards, and all summation categories), all values were regressed toward the mean to make this mechanism comparable to the regressed-frequency mechanism.

According to the risk-category mechanism, the choice between two risks is based on the average frequency in Category *A* (to which *a* belongs) and Category *B* (to which *b* belongs). The prediction can therefore be stated as follows:

$$\text{Choice proportion}_{\text{Risk } a} = \text{Regressed average frequency}_{\text{Category } A}/(\text{Regressed average frequency}_{\text{Category } A} + \text{Regressed average frequency}_{\text{Category } B}),$$

where $\text{Regressed average frequency}_{\text{Category } A}$ and $\text{Regressed average frequency}_{\text{Category } B}$ are the regressed actual average frequencies (i.e., mortality rate or disease incidence) in Risk Category *A* and Risk Category *B,* respectively. Note that the risk-category mechanism predicts that participants are not able to reliably distinguish events from the same category of risks. Consequently, it predicts chance performance in the cancer set and the infection set because they involve within-category comparisons only (e.g., lung cancer vs. breast cancer or syphilis vs. gonorrhea).

Before I turn to Study 4, one comment is in order. One might argue that the availability-by-recall and the fluency mechanisms are at a disadvantage by not relying on regressed values, as do the regressed-frequency and the risk-category mechanisms. Indeed, because both the mapping of the subjective value on the response scale (availability by recall) and the process of retrieval of a term (fluency) are not likely to be devoid of random error, regression to the mean can be expected (Dougherty, 2001; Erev, Wallsten, & Budescu, 1994). Therefore, I decided to treat the availability-by-recall and fluency mechanisms analogously to the other mechanisms. The following analyses are based on the regressed values of the recalled data and the number of mentions. Theses values yielded, in general, the most favorable results for the availability-by-recall mechanism and the fluency mechanism across all studies.

Study 4: Which Mechanism Accounts Best for Judgments of Risk Frequencies?

Study 4 pursued two goals. First, I hoped to replicate the results reported by Lichtenstein et al. (1978). On the basis of this replication, I would then examine which of the candidate processes, if any, could predict people's risk frequency judgments in the present study and, by extension, in theirs. Second, I aimed to test whether the same mechanism could also account for inferences in other sets of health risks involving another criterion (i.e., disease incidences in the cancer and infection sets).

Method

Participants and design. One hundred ten students participated in the study, which was conducted at the Max Planck Institute for Human Development, Berlin, Germany. One group of participants (n = 45) was presented with pairs of causes of death and asked to choose the cause that took more lives (per year). Two other groups of participants (n = 30 and n = 35) were presented with pairs of types of cancer and pairs of infectious diseases, respectively, and asked to choose the disease with the higher incidence rate. All people were paid for participating (a flat fee of €10 [$12.56 U.S.]); half of the participants also received performance-contingent payment according to the following scheme: Two to four participants took part in each session. Within these small groups, the person who achieved the highest percentage of correct inferences received an extra payment of €3 ($3.77 U.S.), the person with the lowest number of correct inferences received no extra payment, and for medium performances, people received €1 ([$1.36 U.S.] in groups of four) or €2 ([$2.51 U.S.] in groups of three). The provision of financial incentives did not affect the results.

Materials. Table 3.1.1 lists the risks included in the assorted set, the cancer set, and the infection set. For all three sets, I determined the annual averaged mortality rates (for the assorted set) and the incidence rates (for the two disease sets) across a 5-year period (from 1996 to 2000), using statistics prepared by the Federal Statistical Office of Germany (Statistisches Bundesamt, 2002) and the Robert Koch Institute (Arbeitsgemeinschaft Bevölkerungsbezogener Krebsregister in Deutschland, 1999; Robert Koch Institute, 2001). In the assorted set, mortality rates in Germany were strongly correlated with those reported by Lichtenstein et al. (1978; Pearson correlation = 0.99). From the assorted set, Lichtenstein et al. constructed 106 pairs (see their Table 2: Lichtenstein et al., 1978, pp. 556–557). I examined the same pairs. From the cancer set, 10 types of cancer were drawn randomly and a set of all possible pairs (45) was constructed. I did the same for the infection set. Both the order in which the pairs appeared and the elements within each pair were determined at random. To make sure that participants understood unfamiliar or ambiguous terms, a glossary was included for some events. If possible, medical jargon was replaced (in the infection and cancer sets) with more commonly used terms. I consulted a physician to assure the equivalence of medical and colloquial terms.[43]

Procedure. After an introductory text explaining the relevance of accurate risk judgments for everyday behavior, people read the following instructions:

> You are asked to judge the annual frequency of occurrence of different [causes of death/ types of cancer/infections] in Germany. . . . Each item consists of two different [causes of death/types of cancer/infections]. The question you are to answer is: For

[43] In one instance, however, the choices of words went astray. The term *food poisoning* (*Lebensmittelvergiftung*) was used to refer to botulism. Although botulism is indeed a form of food poisoning, it is only a special form of it. Not surprisingly, participants estimated food-poisoning incidence to be about 1,300 times more frequent than it actually was. I decided to exclude this item from all analyses, thus reducing the number of pairs in the infection set to 36.

which of two events is the [number of deaths/number of new incidents] per year larger?

Participants were presented with the pairs of risks displayed on a computer screen. After they concluded the choice task, half of the participants continued to work on an estimation task (see Study 6, involving a different set of risks). Half of the participants started with the estimation task first. (The order of the tasks turned out to have no effect.) Because, in the assorted set, the mortality rates of seven causes of death were zero (see Table 3.1.1), for this set participants were not forced to make a choice when they thought a pair to be exactly equally frequent (for three pair comparisons of the assorted set, the actual mortality rates were equal). However, it was stressed that they should use the response option *equally frequent* only after careful consideration. It was used in only 2.5% of all choices.

Results

Before I turn to the test of the mechanisms, I describe the obtained choices in more detail. Table 3.1.2 shows the percentage correct in all three sets. On average, participants scored 71.2% correct in the assorted set, thus approximating the 73.7% correct reported by Lichtenstein et al. (1978). Whereas, in the cancer set, mean accuracy was slightly lower (68.2%), it was markedly higher in the infection set (80.6%). Also consistent with Lichtenstein et al. is the observation that participants' scores in each set varied widely, although the variability is more pronounced in the cancer and infection sets than in the assorted set.

Table 3.1.2. Choice accuracy and item difficulty (i.e., median ratio of more frequent to less frequent risk) in the assorted set, the cancer set, and the infection set.

	Study 4			Study 5	
Percentage correct	Assorted (*n*=45)	Cancer (*n*=35)	Infection (*n*=30)	Cancer (*n*=40)	Infection (*n*=40)
M	71.2	68.2	80.6	62.8	62.1
Mdn	72.6	68.9	79.8	63.8	63.6
Range	58.5–78.3	48.9–82.2	55.6–91.7	51.5–72.1	48.2–74.3
SD	4.67	8.60	8.12	5.05	5.68
Item difficulty (*Mdn* ratio)	10.9	3.5	72.4	3.2	37.4

Why did mean accuracy vary so markedly across sets? I suggest that some of the variation in the scores is due to differences in item difficulty. Ceteris paribus, the smaller the distance between Risks *a* and *b,* the more difficult it is, one can assume, to distinguish between them. One can capture the difficulty of an item in terms of the ratio between the more frequent and the less frequent cases. Figure 3.1.2 shows that participants' percentage correct scores were a function of this ratio: The majority of participants decided correctly once the ratio was about 10:1 or larger. Table 3.1.2 also shows that the median ratio tracked

the average scores: The set with the best performance, the infection set, was the set with the highest median ratio and vice versa. Across all three sets, the majority of participants made the correct choice in 83% (152 out of 184) of all pair comparisons.

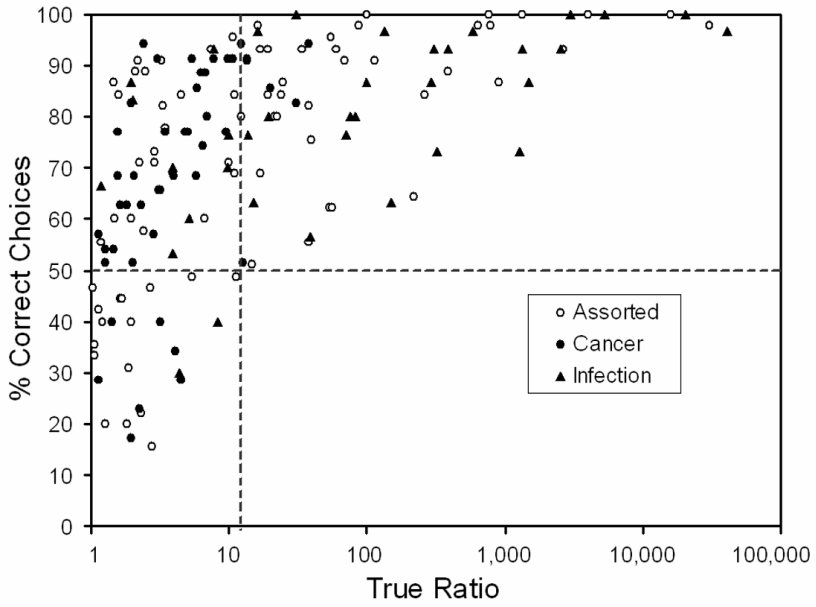

Figure 3.1.2. Choice proportion and item difficulty: percentage of participants who correctly identified the more frequent of two risks as a function of the ratio of more frequent to less frequent risk in the assorted set (empty circles), the cancer set (filled circles), and the infection set (triangles). Twenty-eight of the 106 pair comparisons were excluded from the assorted set because the actual mortality rate of at least one event was zero.

Which mechanism predicted choices best? To answer this question, I used two goodness-of-fit criteria. The first criterion was the distance between actual and predicted choice proportions, measured by root-mean-square deviations (RMSDs). Smaller RMSDs indicate better predictions. Figure 3.1.3 shows the RMSD for each mechanism.[44] Across all three sets, two clear winners emerged. The RMSDs are smallest for the regressed-frequency mechanism and the availability-by-recall mechanism. Except in the cancer set, in which the fluency mechanism performed well, both mechanisms competed markedly better than the fluency mechanism and the risk-category mechanism. The failure of the risk-category mechanism becomes particularly obvious in the cancer and infection sets, which include within-category comparisons only. For such comparisons, the risk-category mechanism predicted that people cannot reliably distinguish between risks. As the level of accuracy reached in both sets testifies (see Table 3.1.2), this prediction is wrong.

[44] Across all four mechanisms, 3 pairs were excluded from the assorted set because their mortality rates turned out to be exactly equally frequent. In addition, for the fluency mechanism, 17 pairs were excluded for which no predictions could be derived (because the terms, e.g., motor vehicle–train collision, did not map onto the way respective events are described in newspaper articles).

Table 3.1.3. Spearman rank correlation coefficients between actual and predicted choice proportions.

	Study 4			Study 5	
Mechanism	Assorted set	Cancer set	Infection set	Cancer set	Infection set
Regressed frequency	.67	.66	.49	.34	.61
Availability by recall	.67	.64	.40	.77	.67
Fluency (media)	.43	.80	-.25	.79	.29
Fluency (speed)				.28	-.11
Risk category	.21	-	-	-	-

Note. For the risk-category mechanism no correlation could be calculated for the cancer and infection sets. Except for the negative, all correlations are statistically significant ($p < .05$; two tailed).

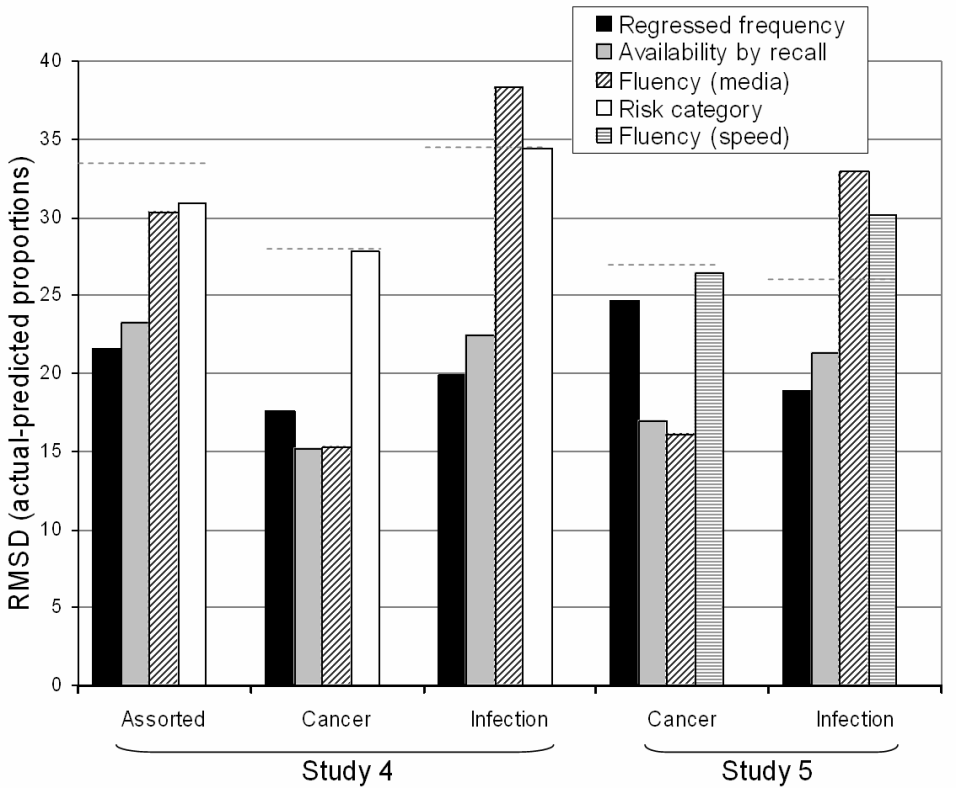

Figure 3.1.3. Which mechanism predicted choices best? Root-mean-square deviations (RMSDs) between predictions derived from the four mechanisms and actual choice proportions in the assorted set, the cancer set, and the infection set of Study 4 and the cancer set and the infection set of Study 5. The dotted lines represent the RMSD level under the assumption of random choice between both risks (per comparison). Note that the risk-category mechanism equals chance performance in the cancer and infection sets.

The RMSD measure does not take into account the pattern predicted by the individual mechanisms. For instance, two mechanisms may have the same RMSD, but one mechanism monotonically follows the data whereas the other zigzags around the data. To quantify the extent to which predictions monotonically followed the data, I computed Spearman rank correlations between predicted and actual choice proportions. As Table 3.1.3 shows, the correlation analysis is consistent with the RMSD analysis: In general, the regressed-frequency

and the availability-by-recall mechanisms competed best and followed the actual data better than the other two mechanisms, except in the cancer set, in which the fluency mechanism performed best.

In sum, I examined which mechanism explained choice data best across various sets of risk. Two criteria of goodness-of-fit—RMSD and Spearman rank correlations between predicted and actual choice proportions—favored the regressed-frequency and the availability-by-recall mechanisms. Although the fluency mechanism fared well in the cancer set, it did not fit the data in the other two sets. Finally, the risk-category mechanism achieved the worst fit across three sets.

Study 5: A Second Test Involving Individual Responses and Another Definition of Fluency

The poor performance of the fluency mechanism in Study 1 was surprising. In line with the common wisdom that media coverage shapes people's risk perception (see also Combs & Slovic, 1979), I counted the frequency of occurrences of words in print media and used such environmental frequencies to define fluency. Of course, this definition of fluency as environmental statistics is only one possible measure of retrieval fluency. Moreover, it could be objected that this measure does not take into account interindividual differences in exposure to occurrences of the terms in the print media. Both of these reasons perhaps caused the inferior performance of the fluency mechanism.

Study 5 was designed to examine the robustness of results of Study 4 by examining an alternative definition of fluency. Specifically, I defined fluency in terms of the speed with which an individual person would recognize the name of, say, a type of cancer or an infection (see also Schooler & Hertwig, 2005). For illustration, readers may notice that when they read the terms *breast cancer* and *hepatic cancer,* they are likely to immediately recognize breast cancer but take a moment to recognize hepatic cancer, if they recognize it at all. The new definition of the fluency mechanism took advantage of this difference in recognition time. It assumed that people could capitalize on such differences in recognition times and that the recognition times would be indicative of the ease with which additional retrieval processes—for instance, bringing instances or occurrences of the event in question to mind—could occur. In the interest of psychological plausibility, however, I assumed limits on people's ability to discriminate between recognition times. Rather than assuming that a person could discriminate between minute differences in any two times, I assumed that if the recognition times of the two risks were less than a just-noticeable difference apart, then the system must guess. Guided by Fraisse's (1984) conclusion on the basis of an extensive literature review that durations of less than 100 ms are perceived as instantaneous, the just-noticeable difference was set to 100 ms (see also Schooler & Hertwig, 2005). It is not claimed, however, that this value captured people's actual thresholds exactly.

A desirable side effect of this definition of fluency was that the mechanism could now also be tested against individual responses. Specifically, the fluency mechanism assumes that if a person recognizes the name of one of two diseases more quickly, then he or she can infer that this disease has a higher incidence rate. To exploit this potential for tests of individual responses, I also derived individual-specific predictions for the other mechanisms: In the case of the availability-by-recall mechanism, I assumed that if a person recalls more instances of one of two diseases among one's social circle, then he or she can infer that this disease also has a higher population incidence rate. For the regressed-frequency mechanism, I assumed that a person retrieves the regressed value of the actual frequencies of both diseases and rests his or her inference on this information. Naturally, in tests of individual responses, the regressed-frequency mechanism is handicapped as it predicts the same choice across all participants for any given pair of diseases.

Study 5 also rendered it possible to examine how robust the good performance of the regressed-frequency and the availability-by-recall mechanisms would be when tested against new samples of items from the same risk environments. In Study 5, I used all 24 elements per set (see Table 3.1.1) and generated all 276 possible pairs per set. In the case of the infection set, this procedure markedly increased the item difficulty (as suggested by the median ratio of more frequent to less frequent risk; see Table 3.1.2). Would the results obtained in Study 4 hold up when mechanisms were tested against these encompassing sets of comparisons?

Method

Participants and design. Eighty students participated in the study, which was conducted at the Max Planck Institute for Human Development. Two groups of participants (each $n = 40$) were presented with pairs of types of cancer and pairs of infectious diseases, respectively. Using the instructions employed in Study 4 (see previous *Method* section), participants were asked to choose the disease with the higher incidence rate. Half of the participants in the cancer group and the infectious disease group were paid a flat fee of €12 ($15.07 U.S.). The other half received a flat fee of €9 ($11.30 U.S.) and, in addition, performance-contingent payments. They earned 4¢ (5¢ U.S.) for each correct answer and lost 4¢ for each wrong answer. As in Study 1, the provision of performance-contingent payment did not have an effect.

Materials. Both the order in which the 276 pairs of types of either cancer or infections appeared and the elements within each pair were determined at random. The assorted set was not included because preliminary tests revealed that some rather long terms (e.g., *motor vehicle—train collision, poisoning by solid or fluid,* and *pregnancy, childbirth, and abortion*) and some rather short terms (e.g., *flood, lightning*) produced extremely uneven response times, thus making a stringent test of the new fluency mechanism difficult. As it had fared badly in Study 1, I did not examine the risk-category mechanism.

Procedure. Prior to their choices, participants were presented with the 24 types of either cancer or infectious diseases (see Table 3.1.1) on a computer screen. The names of the

diseases were presented in random order and one at a time. Participants were asked to decide whether they had heard of this type of cancer or infectious disease before and to express their positive or negative answer by pressing one of two keys. They were instructed to keep the index fingers of the right and the left hands positioned on the *yes* and *no* keys, respectively, for the entire duration of this task and were encouraged to respond as quickly and accurately as possible. The time that elapsed between the presentation of the name and their keystroke was measured. Note that the recognition judgments were collected prior to the choices in order to avoid that the reverse order conflates the recognition judgments. Of course, asking for recognition judgments at the outset may have primed people to rely on recognition or lack thereof in the choice task. This possibility, however, was deemed less problematic because it would work in favor of the fluency mechanism, and Study 5's goal was to give the fluency mechanism a second chance. Finally, as in Study 4, after having completed the choice task, participants indicated for each of the types of cancers or infectious diseases the number of instances they could recall from their social network.

Results

Before I turn to the test of the mechanisms, I first describe the obtained choices in more detail. On average, participants scored 62.8% and 62.1% correct in the cancer and infection sets, respectively (see Table 3.1.2 for more detailed information). The level of accuracy in the infection set was lower than that achieved in Study 4 (62.1% vs. 80.6%). Item difficulty, measured in terms of the ratio between the more frequent and the less frequent risk elements, provides a partial explanation for the decline in accuracy: On average, pair comparisons in the infection set were markedly more difficult in Study 5 than in Study 4 (37.4 vs. 72.4; see Table 3.1.2).

Which mechanism predicted individual choices best? Figure 3.1.4 plots, for each mechanism, how often it rendered possible a prediction per person. Across the total of 552 items (276 items from each set), the availability-by-recall mechanism discriminated on average in only 132 cases (24%); *discriminated* here means that the mechanism arrived at an unambiguous prediction (i.e., predicted either Risk *a* or Risk *b* to be the disease with the higher incidence rate). The low discrimination rate was due to the fact that many participants could not recall any occurrence of the diseases in question within their social circle. Rather than having the mechanism guess, I excluded the respective comparisons from the test set. The fluency and the regressed-frequency mechanisms, in contrast, discriminated on average in 426 (77.1%) and in 552 (100%) cases, respectively. In the case of the fluency mechanism, I included all cases in which one risk was recognized and the other was not, as well as those cases in which both risks were recognized and their respective recognition times differed by at least 100 ms.

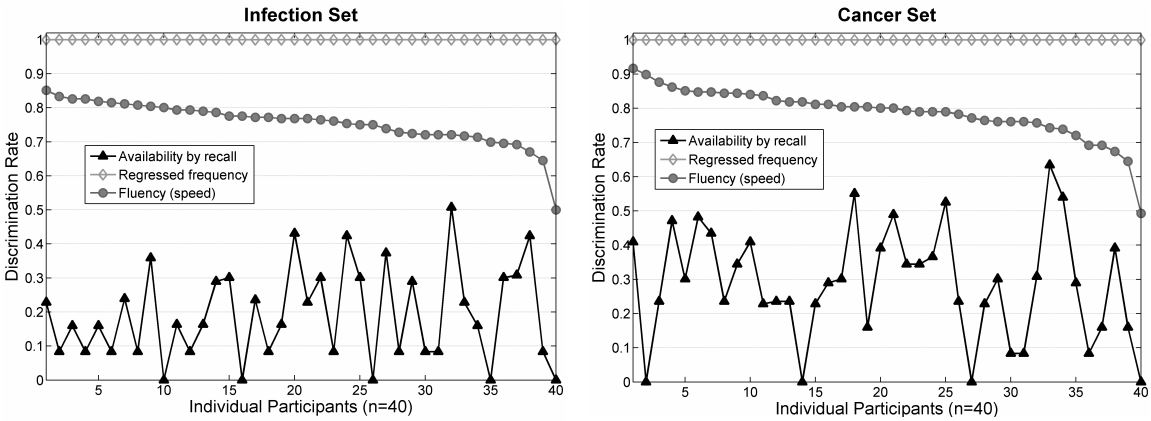

Figure 3.1.4. How often did the mechanisms make a prediction? Discrimination rates (for each of the 40 participants) for the availability-by-recall, regressed-frequency, and fluency mechanisms for the infection set (a) and the cancer set (b).

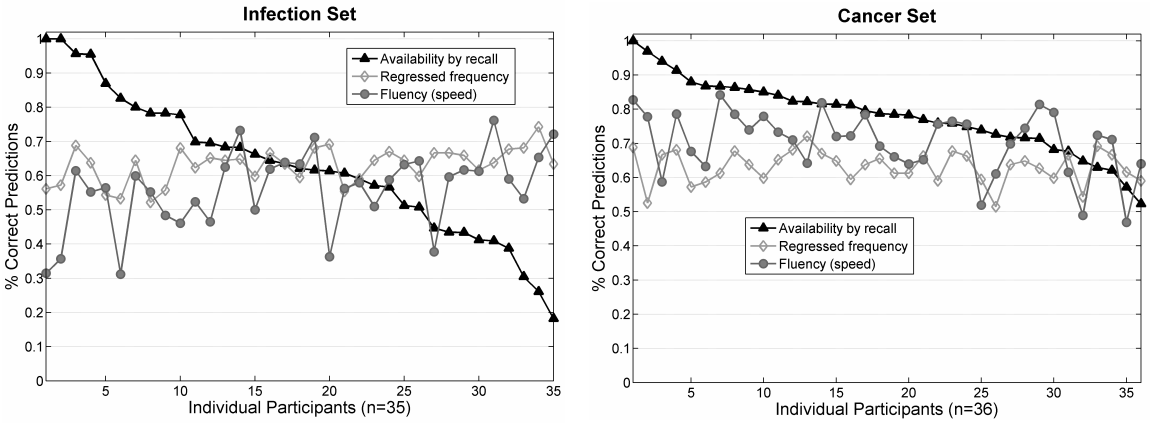

Figure 3.1.5. How often did the mechanisms make the correct prediction? Proportions of correctly predicted actual choices (within the set of comparisons in which a mechanism discriminated) for the infection set (a) and the cancer set (b).

Next, I turn to how often the predicted choice matched the actual choice. Figure 5 plots the percentage of correctly predicted actual choices (within the set of comparisons in which a mechanism discriminated). In the infection set (Figure 3.1.5a), the availability-by-recall and the regressed-frequency mechanisms competed best—62.7% and 62.1% correct predictions, respectively—and predicted the actual choices markedly better than the fluency mechanism (56.6%). In the cancer set (Figure 3.1.5b), in contrast, the availability-by-recall mechanism (78% correct predictions) clearly outperformed the other two mechanisms, whereas the fluency mechanism (69.8%) performed about five percentage points better than the regressed-frequency mechanism (62.8%).

As pointed out, the mechanisms' discrimination rates (see Figure 3.1.4) differed extremely. To level the playing field, I next turned to a different kind of analysis. Specifically, I compared the three mechanisms using critical items. Critical items are pairs in which two

mechanisms discriminate but make a different prediction. For each individual participant, I determined the mechanism that correctly predicted the majority of such critical cases in each of the two contests with the respective competitors. In the cancer set, the availability-by-recall mechanism thus explained 22 participants (out of 31; 9 participants remained unclassified). The fluency and regressed-frequency mechanisms lagged far behind, with 7 and 2 explained participants, respectively. In the infection set, in contrast, the regressed-frequency mechanism explained 17 participants (out of 26; 14 remained unclassified), whereas the availability-by-recall and the fluency mechanisms explained 5 and 4 participants, respectively.

Which mechanism performed best on an aggregate level? Still another way to address the mechanisms' widely different discrimination rates would be to analyze the data on the aggregate level, as in Study 4 (see the Predictions section, above).[45] Such an analysis would have the additional benefit of allowing for a comparison of the results across studies. I used the same goodness-of-fit criteria as in Study 4. As Figure 3.1.3 shows, the RMSDs in the cancer set were smallest for the availability-by-recall and the fluency mechanisms. In the infection set, in contrast, the regressed-frequency mechanism performed best, closely followed by the availability-by-recall mechanism. The fluency mechanism (both definitions) clearly fell behind. The second goodness-of-fit criterion—Spearman rank correlations between predicted and actual choice proportions—corroborated this picture (see Table 3.1.3). Thus, by and large, the analysis on the aggregate level mirrored the results obtained for individual responses.

Summary of Studies 4 and 5

Two studies were conducted with a total of about 30,000 choices. In Study 4, the notion of fluency was defined in terms of number of mentions of a risk in print media. Both criteria of goodness of fit favored the availability-by-recall and the regressed-frequency mechanisms (see Table 3.1.3 and Figure 3.1.3). In Study 5, fluency was defined in terms of the time it took to decide whether one recognized the name of a health risk. In addition, Study 2 tested the mechanisms' predictions against individual responses and against aggregate data. Across the four criteria of goodness of fit—percentage of correct predictions, analysis of critical items, RMSD, and Spearman rank correlation—I found that the availability-by-recall mechanism and the regressed-frequencies mechanism performed equally well in the infection set. In the cancer set, in contrast, availability by recall outperformed the regressed-frequency mechanism and the fluency mechanism (speed) when tested against individual data (see Figure 3.1.5b) and was close to the fluency mechanism (media) when tested on the aggregated level (see Table 3.1.3 and Figure 3.1.3).

On the basis of Studies 4 and 5, I conclude that regardless of whether fluency is defined in terms of word frequency or recognition speed, its predictive power is limited.

[45] Both definitions of fluency were used. For the definition in terms of recognition speed, I used the median recognition time (RT), and the predictions were determined by Choice proportion$_{\text{Risk } a}$ = RT$_{\text{Risk } b}$/(RT$_{\text{Risk } a}$ + RT$_{\text{Risk } b}$) (cf. Sedlmeier et al., 1998).

Across the two studies, different goodness-of-fit criteria, and different test sets, there was a total of 14 contests between the candidate mechanisms. Of these, the fluency mechanism won only 3 of the 14 tests. The availability-by-recall mechanism and the regressed-frequency mechanism each won 5 tests and were tied on 1.[46] This simple counting exercise of tests is admittedly coarse, but the resulting picture is the same for two independent studies: Of the examined mechanisms, the two most promising mechanisms are the availability-by-recall and the regressed-frequency mechanisms.

Study 6: Can the Candidate Mechanisms Also Model Absolute Estimates of Risk Frequencies?

Most people know that, in comparison with most other modes of transportation, it is safer to fly. Yet, to really feel safe, sometimes one would like to know how few people's lives have actually been claimed by plane crashes. Often, such a question comes to mind after one has just buckled oneself into an airplane seat. In this and many other situations, all one can do is to estimate this number. Can the candidate mechanisms account for such absolute estimates of risk frequencies? Applying the four mechanisms to quantitative estimates, however, is not trivial because only two of them lend themselves to predicting absolute quantities: The regressed-frequency mechanism predicts that the estimated number of lives that are taken by, for instance, breast cancer corresponds to the regressed actual mortality rate of breast cancer. The risk-category mechanism predicts that the estimated mortality rate for breast cancer equals the (regressed) average frequency within the category of all diseases. Despite Lichtenstein et al.'s (1978) proposal of the availability heuristic as a possible mechanism for absolute estimates of mortality rates, it does not lend itself directly to predictions of quantitative estimates. One cannot simply take the recalled number of deaths from, say, breast cancer (experienced in one's social circle) as an estimate of the population mortality rate. Instead, one would need to, for instance, estimate how large one's social circle is in relation to the total population and then adjust one's frequency estimates accordingly.

Even without such an intermediate step of extrapolation, however, the availability-by-recall mechanism can be used to predict what Brown and Siegler (1993) referred to as *mapping knowledge.* Mapping knowledge refers to how well people's estimates map onto the ranking of objects according to their actual frequencies. Such a mapping is one property of accurate quantitative estimation. In what follows, I describe how I tested which of the candidate mechanisms could account for mapping properties of frequency estimates.

[46] That is, they both had the same Spearman rank correlation in the assorted set in Study 4 (see Table 3.1.3).

Method

One-hundred sixty-four students participated in the study, which was conducted at the Max Planck Institute for Human Development (these were the same participants who partook in Study 4). Three groups of participants were presented with the assorted set (n = 45), the cancer set (n = 59), and the infection set (n = 60), respectively. Each participant was paid a flat fee of €10 ($12.56 U.S.), and half of the participants also received performance-contingent payment (according to the scheme described in Study 4; instructions explained the concept of mean absolute deviation between predicted and actual frequency and told participants to attempt to minimize this deviation measure). As previously, the provision of financial incentives did not affect the results. Participants were presented with a randomly ordered list of the risks and asked to estimate the annual mortality rate (assorted set) or the incidence rate (cancer set and infection set). To give participants a sense of the frequency metric, they were told that the total number of deaths in a typical year in Germany is around 850,000 (assorted set). Those who judged types of cancer and infections learned that the annual incidence rate in Germany is about 325,000 and 245,000, respectively. As in Study 4, botulism (in the infection set) was excluded from the final analysis.

Results

Before I turn to the candidate mechanisms, let me describe the estimates and their accuracy in more detail. The median estimates for three risk sets are reported in Table 3.1.1 (median estimates as they are not unduly influenced by outliers). Figure 3.1.6 shows the median estimates plotted against the actual frequencies in the assorted set. As did in Lichtenstein et al. (1978), the obtained pattern seems like overestimation of rare risks and underestimation of common risks (but see Hertwig, Pachur, & Kurzenhäuser, 2005, for a discussion of to what extent this pattern is due to unsystematic error variance).

In evaluating the accuracy of quantitative estimates, Brown and Siegler (1993) proposed to distinguish between two components. Mapping knowledge refers to how well the estimates capture the actual ranking of objects. *Metric knowledge,* in contrast, focuses on how well the estimates capture the statistical properties of the frequency distribution of a domain (such as the mean, median, and variance). Knowing such properties helps people to make estimates in the right ballpark. To measure metric knowledge, Brown and Siegler used the order of magnitude error (OME) measure. OME quantifies the discrepancy between true and estimated values and converts the estimation error to a proportion of an order of magnitude (Brown, Cui, & Gordon, 2002; Brown & Siegler, 1993; see also Nickerson, 1981). The absolute OME was computed according to the following formula: $|\log_{10}(\text{estimated value}/\text{true value})|$.

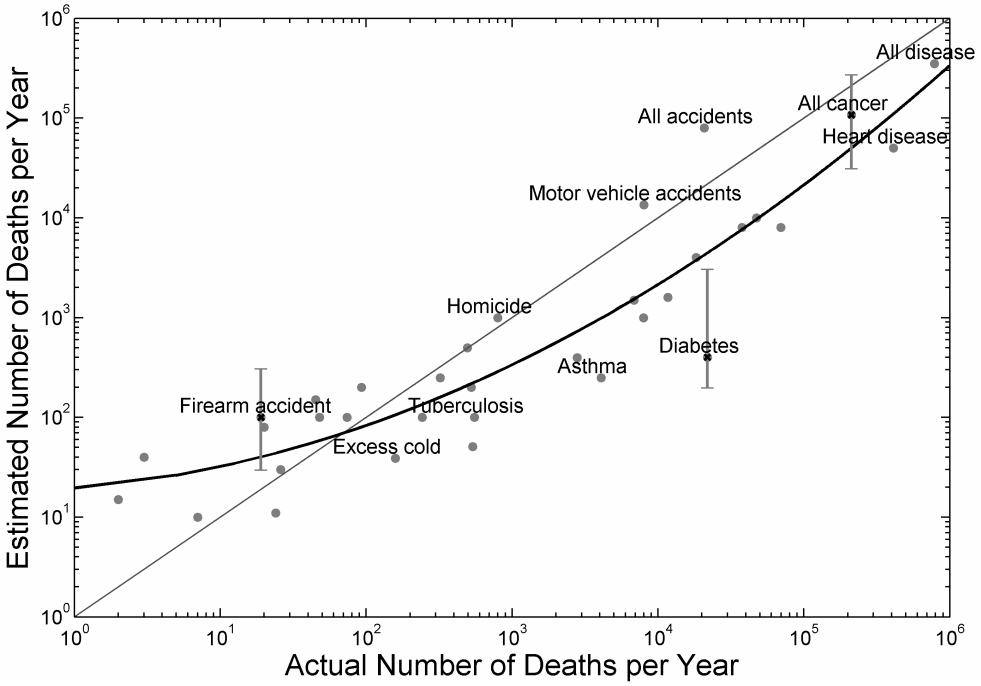

Figure 3.1.6. Estimates of risk frequencies: relationship between estimated and actual number of deaths per year for 41 causes of death in the assorted set. Each point represents the median estimate of 45 participants. The curved line is the best fitting quadratic regression line: Log median = 1.291 + 0.118 × log actual frequency + 0.098 × log actual frequency^2. Vertical bars depict the 25th and 75th percentiles of individual estimates for firearm accidents, diabetes, and all cancer.

Table 3.1.4 reports the mean absolute OME (with standard errors). How appropriate were people's estimates according to this measure? When evaluating the estimates, it is instructive to compare the results that were obtained here with those obtained by Brown et al. (2002). In people's estimates of the population size of 112 nations with 4 million or more people, they found a mean absolute OME of .49. Across all three sets, I found exactly the same mean absolute OME (see Table 3.1.4). This suggests that estimates of health risk frequencies are as accurate as estimates in other knowledge domains. Moreover, I found that accuracy in the infection set was markedly lower in the infection set than in the assorted set and cancer set. Why? One likely reason is that the infection set included numerous very rare events. In fact, a third of all infections have an annual incidence rate of 10 and smaller. Because the incidence rates cannot be negative, people are more likely to err on the high side when estimating the frequencies of infection that are small but constrained to be nonnegative (see also Benjamin et al., 2001). For an infection with an incidence of, say, 1 (e.g., diphtheria; see Table 3.1.1), a deviation of .77 of an order of magnitude would lead to modestly deviating estimates of 5.89 and 0.17 on the high and low sides, respectively.

To evaluate people's mapping knowledge, Brown and Siegler (1993) proposed the Spearman rank correlation. Table 3.1.4 shows these results (the correlation between the median estimate for each risk and its absolute frequency and the median of the individual

participants' correlations). Unlike in the OME measure, the mapping accuracy is comparable in the cancer and infection measures, thus suggesting that how accurate people's estimates are depends on the measure one uses to evaluate them (see also Brown & Siegler, 1993). Across all sets, I found that the median of the individual participants' rank-order correlations is of the same magnitude that Brown et al. (2002) and Brown and Siegler (1993) reported for other domains, namely, around .50, another indication that estimates of health risk frequencies appear not to be different in nature than estimates in other knowledge domains.

Table 3.1.4. Order of magnitude error (*OME*; mean absolute *OME*, standard error), rank correlation between median estimated and actual frequencies (r_s), and the median of the individual rank correlations (and their range) between estimated and actual frequencies.

Accuracy measure	Collapsed (N = 164)	Assorted set (n = 45)	Cancer set (n = 59)	Infection set (n = 60)
M absolute OME	.49	.48	.23	.77
SE	.04	.06	.03	.07
r_s	.86	.93	.55	.63
Mdn individual r_s	.50	.81	.39	.42
Range	-.15–.92	.58–.92	.01–.68	-.15–.75

Which mechanism fit estimates best? The availability-by-recall and the fluency mechanisms render possible predictions regarding the mapping component of estimates but not predictions regarding the metric component. I therefore examined the mechanisms' ability to predict to what extent the estimated values followed the predicted values monotonically.[47] Contrast analysis was used as the measure for the covariation of predictions and estimates (Rosenthal & Rosnow, 1985; Sedlmeier et al., 1998). Table 3.1.5 shows the results of the contrast analysis ($MS_{contrast}$, MS_{error}, df_{error}, and F value).

Table 3.1.5 also shows the effect size r associated with the four mechanisms (Rosenthal & Rosnow, 1985).[48] The larger the (positive) r, the more the data monotonically follow the predictions of the mechanisms. On this measure, the regressed-frequency and the availability-by-recall mechanisms fit the data best across all three sets. The effect sizes for both hypotheses ranged between .74 and .50, corresponding to (very) large effect sizes (J. Cohen, 1988). Although the fluency mechanism fared well in the cancer set (as it did in Studies 4 and 5), it fell behind in the assorted set and the infection set. The risk-category mechanism competed well in the assorted set; however, it could not be tested in the other two

[47] Both definitions of the fluency mechanism were tested, one in terms of environmental frequencies and one in terms of recognition speed. Because it yielded the better results, I report the results only for the environmental frequency definition. To specify the predictions for availability by recall, I computed the sum of the number of recalled instances for each risk across participants in the pilot study (see Prediction section).

[48] The mechanisms' predictions for each individual risk were used to determine the lambda weights, against which people's estimates were contrasted. Weights for contrasts add up to 0. For the calculation of the weights, first, the average of the predictions for a given set and mechanism were calculated. Then, the deviation of the prediction for a single risk from the respective average was used as the weight for that risk. $MS_{contrast}$ (= $SS_{contrast}$, because $df_{contrast}$ is always 1) is calculated as $L^2/n\sum\lambda^2$, where the λs are the derived weights, n is the number of estimates given for each risk, and L is the sum of all weighted (by λ) totals for a given risk.

sets because it would have predicted that within one category, each element would receive the same value (i.e., the weights for contrasts would thus be identical). This prediction would clearly be wrong.

Table 3.1.5. Outcome of the contrast analysis.

Set of risks and mechanism	$MS_{contrast}$	MS_{error}	df_{error}	F	r (effect size)
Assorted set					
Regressed frequency	7,138,608,833,270	46,894,211,008	123.299	152.23	.74
Availability by recall	6,813,457,518,861	46,894,211,008	123.299	145.29	.74
Fluency (media)	1,453,400,797,263	46,894,211,008	123.299	30.99	.45
Risk category	6,894,559,098,174	46,894,211,008	123.299	147.02	.74
Cancer set					
Regressed frequency	72,594,923,710	547,497,314	396.85	132.59	.50
Availability by recall	78,736,346,647	547,497,314	396.85	143.81	.52
Fluency (media)	94,283,416,385	547,497,314	396.85	172.21	.55
Infection set					
Regressed frequency	82,373,753,373	758,308,716	196.57	108.63	.60
Availability by recall	84,141,982,299	758,308,716	196.57	110.96	.60
Fluency (media)	9,545,564,615	758,308,716	196.57	12.59	.25

Note. Because within a set of risks each participant gave frequency judgments repeatedly for the different risks within a set and thus contributed several scores, the MS_{error} and df_{error} were determined by a repeated measures ANOVA (instead of a between-groups ANOVA; see Rosenthal & Rosnow, 1985, p. 12). In all three sets (assorted, cancer, infection), Mauchley's test indicated that the assumption of sphericity was violated. Therefore, the corrected values produced by the Greenhouse–Geisser estimate were used, which produced the fraction numbers for the df_{error}. The risk-category mechanism was not tested in the cancer and infection set because it would have predicted that each element within a set receives the same estimate.

As was the case for judgments of which of two risks is more frequent (Studies 4 and 5), the availability-by-recall and the regressed-frequency mechanisms outperformed the fluency and the risk-category mechanisms in accounting for absolute estimates.

General Discussion

In what follows, I describe the main results and discuss their implications.

What We Have Learned

I proposed and tested four mechanisms of judgments of relative and absolute risk frequencies: two versions of the availability heuristic and two versions of the view that event frequencies are directly encoded and that tallies of environmental frequencies can be retrieved as desired. Two of the four mechanisms received little support. The risk-category mechanism, according to which people's knowledge is limited to a sense of the average frequency in the category, failed most undoubtedly: Out of all four mechanisms, it achieved the worst fit in the assorted set. In addition, it severely underestimated the amount of knowledge that people command about frequencies of infections and types of cancer.

The second account that received at best mixed evidence is the fluency mechanism. Although it competed well with the other mechanisms in the cancer set, it fared badly in the assorted and the infection sets (see Table 3.1.5 and Figure 3.1.3). Ease of retrieval—the notion that Tversky and Kahneman (1973) proposed as one interpretation of availability—is not precisely defined. To turn it into a measurable quantity, I linked ease with the notion of fluency. Fluency was measured in two different ways—in terms of environmental statistics (i.e., frequency of mentions in print media) and in terms of recognition speed (i.e., how quickly people were able to assess whether they had heard of the word in question). The two measures are clearly but not perfectly correlated (Spearman rank correlation between mention frequency and median recognition speed was $r = -.42$ and $r = -.47$ in the cancer and infection sets, respectively). Both measures yielded comparatively good results only in the cancer set. By and large, the results across all three studies do not support the ease interpretation of the availability heuristic. Of course, one cannot exclude the possibility that other definitions of ease, such as number of memory traces and resulting memory strength (instantiated in MINERVA-DM; Dougherty, Gettys, & Ogden, 1999), would have fared better. The results obtained here, however, speak against two quite precise and distinct definitions of ease.

Across different sets of risks, different levels of item difficulty, different kinds of inferences, and different levels of judgmental accuracy, people's inferences conformed best to the predictions of the availability-by-recall and the regressed-frequency mechanisms. Indeed, across all 736 pair comparisons of Studies 4 and 5, the RMSDs for the availability-by-recall and regressed-frequency mechanisms were nearly identical, with values (averaged across the sets) of 19.8 and 20.5, respectively. The fluency and the risk-category mechanisms, by comparison, performed clearly worse, with RMSDs of 26.6 and 29.2, respectively.

Similarly, in Study 6, availability by recall and regressed frequency showed the largest effect sizes except in the cancer set (in which the fluency mechanism reached, by a small margin, the highest effect size). One way of directly comparing the two mechanisms would be to quantify their difference by comparing the respective contrast weights (Rosnow & Rosenthal, 1996, p. 256; see also Sedlmeier et al., 1998, footnote 48) across all three sets. This comparison resulted in a weighted (by *df*) mean effect size of $r = .0001$ (see Table 3.1.6). The differences are thus negligible. It seems fair to conclude that the availability-by-recall and the regressed-frequency mechanisms achieve nearly identical predictive accuracy in modeling people's estimates.

However, the fact that their mean accuracy in modeling people's choices and estimates is indistinguishable does not mean that the mechanisms' predictions are indistinguishable. Take, for instance, the correlation between the predictions of the regressed-frequency and the availability-by-recall mechanisms: although, in both Study 4 and Study 5, the correlations are significant in all sets, they are far from perfect, that is, *r*s = .71, .87 (Study 5: .62), and .41 (Study 5: .26) for the assorted set, the infection set, and the cancer set, respectively. Another example refers to the prediction of inaccurate choices. In 199 of the 736 pair comparisons

(27%) of Studies 4 and 5, the majority of participants selected the less frequent event. Using regressed values of the objective frequencies, the regressed-frequency mechanism could not predict choice proportions smaller than 50%; thus, it fared relatively badly in predicting those 199 choices. In contrast, the availability-by-recall mechanism correctly predicted 162 of those 199 items' choice proportions lower than 50%. However, it also predicted choice proportions lower than 50% in 74 pairs in which the actual choice proportion was above 50%. In other words, there are clusters of items favoring the regressed-frequency mechanism, and others favoring the availability-by-recall mechanism. In addition, Figure 3.1.5 shows that the availability-by-recall mechanism predicted the choices of some participants very well (e.g., for 22 participants, it correctly predicted more than 80% of inferences) but failed in explaining others.

I take these findings to suggest that people have a toolbox of different strategies and, in addition, that they can switch back and forth between different kinds of information (Betsch, Siebler, Marz, Hormuth, & Dickenberger, 1999; Brown, 2002a; Payne, Bettman, & Johnson, 1993). Thus, the same person is not likely to use the same mechanism for each single inference. For instance, if a person cannot retrieve any episode within his or her social circle, he or she may attempt to rely on a sense of fluency or frequency. The likely fact that a person has a repertoire of strategies and can discount a previously used dimension of information (Oppenheimer, 2004) may be key to understanding why the fit for any single strategy is far from perfect in the analyses.

Table 3.1.6. Predictive power of the contrasts for the availability-by-recall mechanism relative to those for the regressed frequency mechanism.

			Regressed frequency	
Set	MS_{error}	df_{error}	$MS_{contrast}$	r
Assorted	46,894,211,008	123.299	14,434,049,726	-.0025
Cancer	547,497,314	396.85	218,893,629.8	.001
Infection	1,269,020,591	196.589	55,938,612.24	.0002
Weighted *M* (by *df*)				.0001

Note. New contrasts were created out of the differences between the original contrast weights (see Rosnow & Rosenthal, 1996). Results are based on the estimation task of Study 6. The F value can be calculated by dividing $MS_{contrast}$ by MS_{error}. The correlation coefficient r as a measure of effect size is calculated by the formula $r = [F/(F + df_{error})]^{1/2}$ (e.g., Rosenthal & Rosnow, 1991).

Retrieving Episodes From One's Social Circle: An Ecologically Valid Cue

Two seemingly quite dissimilar mechanisms conform best to people's judgments of relative and absolute risk frequencies. The availability-by-recall mechanism assumes that people draw samples of the events in question and then use the sample frequencies to estimate the criterion. In contrast, the regressed-frequency mechanism assumes that people automatically encode event frequencies and thus are able to produce accurate (albeit regressed) judgments of relative and absolute risk frequencies. That the two mechanisms are close competitors in explaining people's judgments is surprising: Whereas the latter ascribes

knowledge of actual (regressed) frequencies to people, the former has typically been invoked to explain inaccurate judgments.

Indeed, I am not aware of a single experimental or theoretical attempt to demonstrate how the availability heuristic enables successful inferences. This need not have been so. In their initial framing of the availability heuristic, Tversky and Kahneman (1973) stressed that "availability is an ecologically valid clue for the judgment of frequency because, in general, frequent events are easier to recall or imagine than infrequent ones" (p. 209). That the frequency of recalled instances can be a valid cue for the actual frequencies is exactly what was obtained: The Pearson correlations (Spearman rank correlations) between the number of recalled cases and their actual frequencies in Study 4 were r = .87 (.77), r = .72 (.61), and r = .66 (.43) in the assorted set, the cancer set, and the infection set, respectively; in Study 5, the respective correlations were r = .59 (.46) and r = .98 (.36) in the cancer set and infection set, respectively.

Why is the recalled content a relatively valid predictor for the actual frequencies even though availability is often equated with biased frequency judgments? I suggest that one reason is the space in memory that the availability-by-recall mechanism can search. By requiring participants to recall personally experienced instances of death and illness, the search space was defined as that of the social circle of a person, that is, his or her family, friends, and acquaintances. In contrast, those who have argued that distortions in estimates of risk frequencies are caused by media coverage seemed to assume that the search space in memory extends far beyond a person's social circle and includes a *virtual circle,* that is, his or her encounters with death and diseases that are conveyed through mass media (e.g., Lichtenstein et al., 1978). In fact, had people searched in their virtual circle and used this information as a proxy for the actual frequencies, their estimates would more likely have been distorted. The frequency of mentions in print media is a poorer predictor for actual frequency than are the recall data: The Pearson correlations (Spearman rank correlations) between the number of mentions and the actual frequencies were r = .43 (.74), r = .59 (.44), and r = .21 (.23) in the assorted set, the cancer set, and the infection set, respectively (see also Burger, 1984; Combs & Slovic, 1979; Frost et al., 1997; and Kristiansen, 1983).

Clearly, augmenting the search space in memory by one's virtual circle comes at the price of systematic error. Because of fierce competition for patronage, potential news items are screened for their ability to captivate an audience; thus, the media focus on and amplify certain aspects of reality while scaling down others (Meyer, 1990). As a consequence, event frequencies in the virtual world and the real world can systematically diverge. Thus, if one samples from the virtual world, one would likely arrive at sample statistics that deviate from population statistics. It is, however, not the sampling process that is distorted but the reference class from which one samples.[49] In contrast, sampling within one's social circle guards

[49] This is different from other illustrations of availability in which the sampling process itself is biased. In the letter study, Kahneman and Tversky (1973) assumed that the process of sampling exemplars, that is, words with the letter *r* in the first and the third positions, is distorted because it is more difficult to retrieve words with *r* in

against the media's selection of rare, vivid, dramatic, emotional, and sensational events. Fortunately, in a person's limited social circle, death is sufficiently rare and dramatic that, in all likelihood, each instance would be retrieved regardless of whether a family member died in a plane crash or from a heart attack.

Conclusion

If indeed humankind is about to enter the age of new plagues, in which factors such as overpopulation, poverty, and global climate change pave the way for new health risks, it becomes even more important to better understand how the public perceives and judges risks. The public's perception plays a key role in the political discourse about how a society ought to respond to emerging risks to public health and well-being—as the global debates on how to respond to the risk of terror or new viral illnesses such as SARS amply demonstrate. The investigations reported here should be seen as another step toward developing more precise models of the cognitive underpinning of inferences about the environmental statistics of risks.

---

the third position. It is interesting to note that Sedlmeier et al. (1998, pp. 756–758) found little evidence for this assumption.

3.2 Cues or Instances: What is used for Inferences about Event Frequencies?

According to Jerome Bruner, "the most characteristic thing about mental life" is that we constantly "go beyond the information given …, fill in gaps, … extrapolate" (1972, p. 218; p. 237). We are able to make such inferences by exploiting redundancy in the environment, that is, by using available information that is correlated with the unknown property. Often this information has to be retrieved from memory, underlining the intimate, but often overlooked, relationship between memory and decision making (Hastie & Park, 1986; Dougherty, Gronlund, & Gettys, 2002; Weber, Goldstein, & Barlas, 1995). What types of memory do we use for inferences about the environment? A popular distinction between different forms of knowledge in permanent memory was proposed by Endel Tulving (1972, 1983; for a recent review, see 2002). He distinguished between semantic memory, referring to general, encyclopedic knowledge of the world (e.g., that basketball is a ball sport), and episodic memory, memory of autobiographical, personal experiences that can be located at a specific time and place (e.g., my playing basketball last Sunday). Although the general validity of this distinction has been questioned (e.g., McKoon, Ratcliff, & Dell, 1986), its heuristic value seems to be generally accepted. Might it also be useful for distinguishing between mechanisms of judgment and decision making?[50]

In particular, I am concerned with inferences about the frequency with which events occur in the environment. For instance, which disease is more common in a population: breast cancer or skin cancer? The judgment and decision making literature proposes two approaches to such inferences, which roughly map to the use of semantic and episodic knowledge. On the one hand, one could take advantage of knowledge about general features of the target objects—semantic knowledge—that are correlated with the unknown feature (e.g., Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, Jones, Olsson, & Winman, 2003), in this case frequency in the population. Accordingly, one could consider general features of the diseases such as their deadliness or their possible causes, which might be correlated with the frequency of a disease and use these as cues for an inference. A popular way to represent such knowledge of features mathematically is as a column vector, containing in its rows the event's values on the features (e.g., Clark & Gronlund, 1996). I refer to inferences based on such general features as *cue-based*.

Alternatively, one could access episodes, that is, memory of personal experiences with occurrences of the diseases for an inference. As these occurrences represent a sample of the population, they reflect characteristics of the population (e.g., "global frequencies"). In other

[50] It was not assumed, of course, that semantic knowledge is not involved at all when episodic knowledge is used (e.g., for identification of sport semantic knowledge is required, which is completely in line with Tulving's original conception; Tulving, 2002). Rather, the question is whether semantic or episodic knowledge is the primary knowledge base used for an inference.

words, to infer which of the two diseases occurs more frequently in the population the number of instances of these diseases one can retrieve are used (Chapter 3.1; Tversky & Kahneman, 1973, 1974; Benjamin, Dougan, & Buschena, 2001). Note that being learned by personal experience, such idiosyncratic "local frequency knowledge" constitutes episodic knowledge. Memory of such episodes is often represented mathematically in terms of traces (each represented by a vector; e.g., Hintzman, 1988; Dougherty, Gettys, & Ogden, 1999). I refer to inferences based on such episodes as *instance-based*.

The vector-based memory representation invites a distinction between the cue-based and instance-based approaches in terms of the direction in which information is considered for an inference. Using Fiedler's (1996) terminology, the cue-based approach describes an *intensional* search of information within a vector representing the event type, whereas the instance-based approach describes an *extensional* search for information across traces of occurrences of the event.

Which approach—cue-based or instance-based, relying on semantic and episodic knowledge, respectively (for similar distinctions, see Robinson & Clore, 2002; Juslin et al., 2003)—provides a better description of how people make inferences about event frequencies? The aim of this chapter is to directly compare these two approaches. In particular, I am concerned with inferences about the frequency of events of which people are manifestations: people contract diseases, people have names, and people engage in different kinds of hobbies.

Note that, in contrast to the substantial work on frequency judgments in cognitive psychology (e.g., Watkins & LeCompte, 1991; Williams & Durso, 1986), which typically assume that the judgments are based on some operation (actual recall, fluency, or automaticity) resulting from the encounters with critical events, the judgment task examined here is not concerned with frequencies that people have been exposed to completely. As the to-be-judged frequencies concern frequencies in the population—and a person will typically have been exposed at best to a sample of the population—the frequencies of the events have to be inferred based on information only probabilistically related to the criterion.

The remainder of this chapter is organized as follows. I start by sketching the two approaches to inferences about frequency and describe various candidate mechanisms for each of the two approaches. In Study 7, I develop and test—both in computer simulations and an empirical study—a new heuristic for instance-based inferences about frequency. In Study 2, I test this heuristic against competitor mechanisms and, more generally, contrast cue-based and instance-based mechanisms for frequency judgments.

Two Ways to Infer Distal Frequencies

How could one proceed to decide which of two events occurs more frequently in the population? An extremely simple strategy applies when the name of one event is recognized but not the other. According to the recognition principle (Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 2002), instantiated in the recognition heuristic, the recognized event

is simply inferred to be more frequent. An increasing body of research supports the notion that—especially under limited knowledge und limited time—people rely on the recognition heuristic to judge quantities in the world (Chapter 2; Goldstein & Gigerenzer, 2002; Reimer & Katsikopoulos, 2004; but see Newell & Shanks, 2004; Oppenheimer, 2003).

However, if the names of both events are recognized more information has to be gathered (unless a decision is made by guessing). In the following, I describe two different approaches that apply in this situation, involving either the retrieval of probabilistic cues from one's semantic knowledge or the retrieval of instances of the events from one's social environment.

Frequency judgments based on semantic knowledge: Cue-based mechanisms

According to the standard approach in models of human inference, knowledge about general characteristics of objects (or events) is used to infer unknown properties of them. Such an approach is at the heart of Brunswik's idea of vicarious functioning and often successful as characteristics of objects tend to be intercorrelated (cf. Brunswik, 1955). Applied to inferences about event frequencies, events that are frequent in a population often have general characteristics, or features, that rare events do not have. For instance, a profession that requires an extremely long training (such as medical doctor) might be less likely to be taken up—and thus are rarer—than one that can be performed after only a short training (e.g., car mechanic). Such cues, which represent semantic knowledge as they refer to general properties of the events, might also be used to make an inference about how often a given event occurs in the population.

The notion that an unknown target variable is inferred by using probabilistic cues is a basic assumption in a number of models of human judgment. For instance, in Gigerenzer et al.'s (1991) theory of Probabilistic Mental Models, an inference of which of two objects has a higher criterion value is made by retrieving features from long term memory that are correlated with the judgment objects (e.g., that a city is a state capital). Similarly, Hammond's work on clinical judgment describes human judgment as an integration process of features of the patient (Chapter 1; Hammond, 1955; Hammond et al., 1975).

So far, cue-based mechanisms have received only little attention in the frequency judgment literature. For instance, they are not considered in Brown's (2002a) Multiple Strategy Perspective framework on different ways to judge frequency.[51] One possible reason for this neglect is that rarely a distinction is made between, on the one hand, judgment tasks in which the number of directly experienced instances represents also the entity to which the judgment refers (e.g., Tversky and Kahneman's, 1973, famous names study) and judgments tasks in which the number of directly experienced instances is only a sample of this entity

[51] Admittedly, Brown (2002) considers a "nonnumerical nonenumeration strategies" that are based on "a fact or impression from memory that expresses frequency relevant information in a nonumerical manner" (p. 47). Although this could be extended to include semantic facts that are predictive, and thus indirect indicators, of event frequency, Brown only refers to vague quantifiers (such as "a lot" and "many") that are direct indicators of frequency.

(e.g., Lichtenstein et al.'s, 1978, risk frequency study). A perfect reconstruction of all experienced instances always leads to a correct judgment in the former task, whereas this is not (necessarily) the case in the latter one. Due to the potential of retrieving instances cues will probably be little used when all relevant instances have been experienced directly, whereas the idiosyncrasies of a personal sample of instances might discredit instances as a basis for an inference when only a sample of the relevant population has been experienced. Cues might then come into play also because they reflect general properties (such as population frequency) better than samples, whose constitution varies from person to person.

To make an inference (e.g., which of two events in more frequent) based on a set of cues, the cues can be processed in different ways. I describe three candidate cue-based mechanisms. Cues can have positive cue values, indicating a higher criterion value, and negative cue values, indicating a lower criterion value. The cues differ in their validity, which is defined as the conditional probability of making a correct inference under the condition that the cue allows for an unambiguous prediction (i.e., discriminates between the events). The validity v_i of a cue i is determined by $v_i = R / (R + W)$, where R (W) is the number of correct (incorrect) inferences the cue makes (cf. Martignon & Hoffrage, 2002; Gigerenzer & Goldstein, 1996). The first cue-based mechanisms is the weighted additive mechanism, which weights each cue depending on its validity and integrates all weighted cue value for each event.

Weighted additive linear mechanism (WADD). The cue values are multiplied by the validity of the corresponding cue and the products summed across all cues for each event. The event with the higher sum is inferred to be more frequent in the population.

A simplified additive mechanism does without differential weighting, apart from the direction of the cue, and weights all cues equally (e.g., Dawes, 1979).

Equal weight linear mechanism (EQW). The equal weight linear mechanism gives all cue values a unit value, with +1 if a cue points to a high criterion value and -1 if the cue points to low criterion value. For each event the cue values are summed up. The event with the higher sum is inferred to be more frequent in the population.

An even simpler cue-based mechanism consists of considering only a subset of the available information.

Take The Best (TTB). Instead of integrating all available cues, the lexicographic inference heuristic Take The Best (Gigerenzer & Goldstein, 1996) searches for cues sequentially according to their validities. As soon as a discriminating cue is encountered, TTB infers that the event to which this cue points is more frequent in the population.

Contrary to the previous cue-based mechanisms, Take The Best is *noncompensatory* as once information search is stopped no amount of evidence represented later in the cue hierarchy can reverse the decision made after search is stopped. There are various studies showing that people use such one-reason decision making strategies, in particular when information costs matter (e.g., time pressure, inferences from memory, external information

search costs; Bröder, 2000; Bröder & Schiffer, 2003a; Newell & Shanks, 2003, Rieskamp & Hoffrage, 1999).[52]

Frequency judgments based on episodic knowledge: Instance-based mechanisms

Events that occur frequently in the population are also more likely to be encountered in one's limited social sphere. Therefore, one could take instances one personally knows as "keys to assessing the distal environment" (Fiedler, 2000, p. 661), that is, as an indicator of the events' overall frequencies. Such an approach has been instantiated in multiple-trace memory models such as Minerva-DM (Dougherty et al., 1999) and BIAS (Fiedler, 1996). Tversky and Kahneman (1974) described such a strategy as an example of the availability heuristic (see p. 1127), although the availability heuristic is also compatible with other mechanisms (e.g., Betsch & Pohl, 2002). To avoid this ambiguity, I specifically refer to the reliance on instances retrieved from one's social network as the *recall principle*.

How are instances in one's social environment processed when used for inferences about event frequencies in the population? One possibility is that to make an inference all instances in a person's social network are retrieved. In Chapter 3.1 I called such a mechanism *availability by recall* and found that it was able to predict people's judgments of risk frequencies robustly across different risk domains and task formats.

Availability by recall (RECALL). According to this mechanism, an inference is based on the total number of instances recalled from a person's social network (across self, family, friends, and acquaintances) for the events in question. The event for which a higher number of instances was retrieved is inferred to be more frequent in the population.

But is it plausible to assume that people will always retrieve all instances they know? For instance, in Epstein's (2001) model of norm generation individuals are assumed to sample only a very limited number of network members. Similarly, Tversky and Kahneman (1971) argued that people often make decisions based on only a small number of observations. As Gigerenzer et al.'s (1999) fast and frugal heuristics often consider only a subset of cues for a decision, it is possible that people retrieve only a subset of instances. But if only a limited number of instances are retrieved from memory, when does one stop sampling from one's social circles?

I propose a new mechanism, the *social circle heuristic*, which assumes that the structure of a person's social network is used to guide and stop the sampling process. Accordingly, it is assumed that an individual's social network has a hierarchical structure, with the relationships that a person has to the members of her social environment differing in genetic relatedness, frequency of contact, emotional closeness, and function of contact (e.g., Hill & Dunbar, 2003; Milardo, 1992; Zhou, Sornette, Hill, & Dunbar, 2005). A popular notion in social network research has been to represent the hierarchical structure of a social

[52] Interestingly, both EQW and Take The Best have been found to be able to outperform more complex mechanisms such as multiple regression (Czerlinski, Gigerenzer, & Goldstein, 1999; Dawes & Corrigan, 1974; Gigerenzer & Goldstein, 1996), as they are less likely to fit noise in the data.

network in terms of concentric circles of varying radius (Kahn & Antonucci, 1980; Moreno, 1936). The social circle heuristic works by sequentially sampling instances of the events in question from the different circles, starting with the focal circle (which is oneself).

According to the social circle heuristic each circle is considered sequentially, starting with the focal circle. As soon as from a given circle more instances can be retrieved for one event compared to the other event, search is stopped and the event for which more instances could be retrieved is inferred to be more frequent in the population. The heuristic is shown in the form of a flow diagram in Figure 3.2.1.

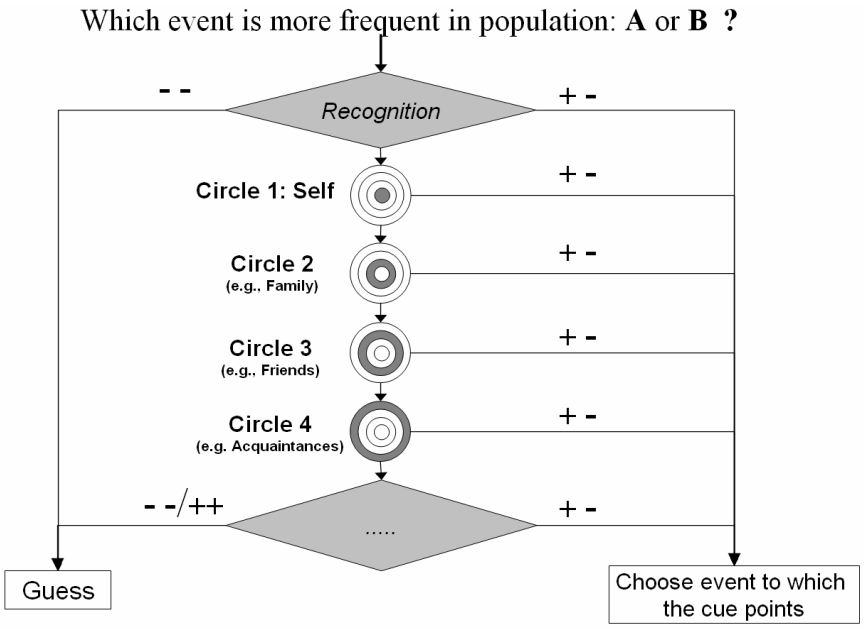

Figure 3.2.1. Flow diagram of the social circle heuristic (here as the SCH-Group version) and the relationship of the sampling process to the recognition principle and inferences based on other cues (such as Take The Best; Gigerenzer & Goldstein, 1996).

Since the heuristic makes a decision as soon as a circle allows for an inference, the search process will often be terminated early, and an inference will be derived from samples of small sizes. Similar as TTB, the heuristic is noncompensatory, since if an inference is made, the information of later circles cannot overturn the decision anymore. The notion of a sampling process of instances guided by social circles proposed by the heuristic emerges from a memory perspective: it could be argued that the circles, demarcating groups of different emotional closeness, length of acquaintance or frequency of contact, are a proxy for how fluently instances in these groups can be retrieved. For instance, information about people one feels very close to or who are contacted frequently should also be retrieved more fluently than

information about people contacted rarely, as the memory traces are stronger (e.g., Anderson & Schooler, 1991).[53] How could the circles be defined? Two versions are considered.

Social circle heuristic-Group (*SCH-Group*). In the first version, the circles are defined by the functional group, of which three are distinguished here: family, friends and acquaintances (cf. Dunbar, 1996). The innermost circle (Circle 1) thus represents the person herself, the next circle (Circle 2) represents the person's immediate family, that is, relatives. The third circle represents the person's friends and the outer circle (Circle 4), finally, represents the person's acquaintances, that is, members in the person's social network which the person knows only superficially.

Another way to differentiate members of one's social network is in terms of how often one typically has contact with them (Hill & Dunbar, 2003). Therefore, I consider a second version of the social circle heuristic in which the circles are defined by frequency of contact.

Social circle heuristic-Contact (*SCH-Contact*). Circle 1 (self) is the same as in SCH-Group, but the subsequent circles are defined by frequency of contact rather than social group. Three further groups are differentiated. The first includes those members contacted at least once a week (Circle 2), the second group those contacted once a month (Circle 3), and the last one those contacted once in 6 months or less (Circle 4). As in SCH-Group, instances are assumed to be retrieved sequentially from the circles and an inference is made as soon as the number of instances retrieved from one circle discriminates between the events.

General Overview of Studies

In a first step, I examined the instance-based approach. First, I explored whether inferences that are made on the basis of a subsample of all instances in people's social networks can produce accurate inferences? Second, is the instance-based approach able to predict people's inferences about frequencies correctly, and which of the three mechanisms is most suitable? Study 7 includes a computer simulation and an empirical study and focuses on the comparison between the social circle heuristic, with circles defined by groups (SCH-Group), and availability by recall.

In Study 8, the instance-based approach is tested against the cue-based approach for predicting participants' inferences. In addition, it is explored whether people's inference processes are adaptive, that is, whether people select those mechanisms that reach the highest accuracy for a particular environment, as can be argued, for instance, from the "adaptive decision making" view (Payne, Bettman, & Johnson, 1988; 1993) or by the ecological rationality view (Gigerenzer, Todd, & the ABC Research Group, 1999).

[53] Importantly, compared to cue-based mechanisms, it is not the validity of information that determines the order in which evidence is considered.

Study 7

To test the accuracy of the social circle heuristic relative to the less frugal availability by recall, I conducted a computer simulation where the task was to infer which of two events, A or B, occurs more frequently in the entire population. For this task, the heuristic could search for instances of the events in its spatial vicinity. I created a population consisting of 2,500 agents, represented in a 50 × 50 grid, in which each cell represented one agent (see Figure 3.2.2a, which shows the environment simplified to a population with 100 agents in a 10 × 10 grid). In the environment, instances of 10 events were distributed randomly across the 2,500 agents (see Figure 3.2.2b). The 10 events mimicked the frequency distribution of occurrences of infectious diseases in Germany, which were also used as the environment in the empirical study (see below). As can be seen from Figure 3.2.2b, the distribution of the proportions of the diseases is highly skewed and falls into a J-shaped distribution, a pattern found in many real world domains (Hertwig, Hoffrage, & Martignon, 1999). The proportions of the 10 most frequent infectious diseases (from a set of 24) were chosen because their proportional distribution could be represented in a population of 2,500 agents. The most frequent event was set at a frequency of 2,000; the 9 other events were distributed according to this anchor and the proportions reported by the Robert Koch Institute.

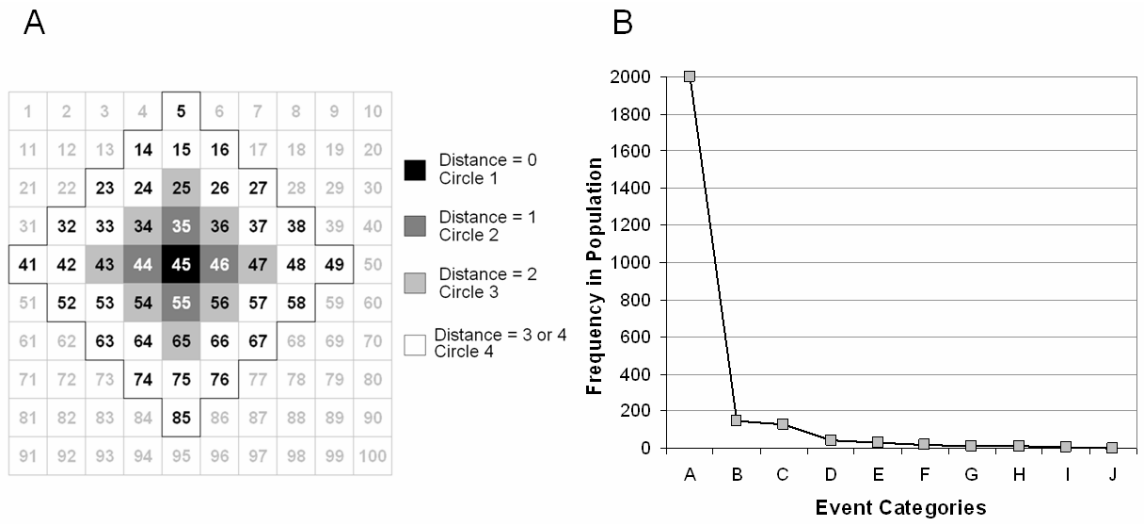

Figure 3.2.2. (a) Representation of the population in the computer simulation (here simplified as a 10 × 10 population). (b) Frequency distribution of the 10 events in the computer simulation.

The social networks were conceptualized according to the distance between the agents, defined by a city-block metric. For instance, in Figure 3.2.2a—which shows the social network of agent #45 in a population of 100—agent #44 has a distance of 1 from agent #45, agent #34 has a distance of 2 from agent #45, agent #33 has a distance of 3 from agent #45 etc. It is assumed that each agent's social network consists of 40 other agents that differ with regard to their distances to the agent. Thus, an agent could maximally sample information

about 41 agents (including himself). This social network was divided into four different social circles: Circle 1, that only included the agent itself, Circle 2, including all neighboring agents with a distance of 1 (four agents), Circle 3, including all neighboring agents with a distance of 2 (eight agents), and Circle 4, including all neighboring agents with a distance of 3 or 4 (28 agents).

To make an inference, the social circle heuristic started with Circle 1 and looked whether event A or B was present. If one was and the other not, no further circles were looked up, and it was inferred that the retrieved event was more frequent in the population, otherwise the next circle was considered, and so on. In case no circle discriminated one of the events was picked randomly. In contrast, availability by recall retrieved all relevant instances in the social network. The event for which more instances could be retrieved was inferred to be more frequent in the population. In case an equal number of instances was retrieved for both events, or if no instances could be retrieved at all, one of the events was picked randomly.

The simulation was repeated 100 times, such that for each run of the simulation the instances of the 10 events (totalling around 2,400 instances, that is, not all agents were instances of one of the 10 event categories) were randomly distributed, and each time 100 agents were picked randomly as starting points to determine the accuracy for both inference mechanisms. At each run, the 10 events were combined in a complete pair comparison (yielding 45 pairs) and the task was to infer which event is more frequent in the entire population.

How well did the social circle heuristic perform compared to availability by recall? Surprisingly, both mechanisms showed an identical proportion of correct inferences with a median of 77.8% (social circle heuristic: M = 76.3%, SD = 7.1, availability by recall: M = 77.5%, SD = 7.1). Showing a similar level of accuracy, the social circle heuristic sampled, on average, only 4.7 instances, which is only around half the amount of information that availability by recall used, which sampled on average 7.9 instances.

In sum, the computer simulation shows that the social circle heuristic can compete with availability by recall, although it uses only a subset of all information. But do people use such a simple strategy for making inferences about event frequencies? To find out, I examined how real people solve the task of judging the frequency of diseases.

Method

Participants. Forty students (27 female and 13 male, mean age 24.2, range 18-31) from various subject fields participated at this study. It was conducted at the Max Planck Institute for Human Development in Berlin. Half of the participants received their payment, in part, depending on their accuracy in the choice task (see below): they obtained an initial fee of €9 (= $11.76 U.S.) and an extra 4¢ (= 5¢ U.S.) for every correct decision; 4¢ were subtracted for every incorrect decision. The other half of participants were paid a flat fee of €10 (= $13.07 U.S.).

Materials. The 24 infectious diseases (the proportions of the 10 most frequent of these were also used in the computer simulation) for which official records are kept by the Robert Koch Institute (e.g., Robert Koch Institute, 2001; see Chapter 3.1) were combined in a complete paired comparison, yielding 276 pairs. In the *choice task* participants were asked to pick the infectious disease that has a higher annual incidence rate in Germany. After this test, participants indicated in a *recall task* for each disease and each of their social circles (self, family, friends, and acquaintances) how many, if any, people had been affected by the disease. They also indicated whether they recognized the name of the disease (*recognition task*). From this information, I calculated how often the participants had an opportunity to choose in accordance with the social circle heuristic and determined which prediction the social circle heuristic made in each of these cases (only comparisons where the names of both names were recognized and the reported number of instances in the social network discriminated between the two diseases were included). The total number of instances for each disease was used to determine the predictions of availability by recall.

Procedure. The choice task always preceded the recall task whereas the order of the choice and the recognition tasks was counterbalanced. For the choice task, participants were seated in front of a computer and received written and verbal instructions for the tasks. Participants saw the 276 pairs of infectious diseases sequentially on a computer screen (presented in 12 blocks) and were instructed to indicate (by pressing one of two response buttons) for which of the two diseases they thought there is a higher number of new cases per year in Germany. The order in which the diseases appeared within a pair was random, as was the order in which pairs was presented, with each participant receiving an individual random order. The recognition and the recall tasks were administered as paper and pencil questionnaires. Each session took around 60 minutes.

Results

Neither order of choice and recognition tasks nor incentives had an effect on the relevant dependent variables. On average, participants made 60.9% (SD = 5.6) correct choices. Overall, only a relatively small number of instances of the diseases were reported by the participants, with an average of 4.2 (SD = 4.7) instances per participant. Due to this low number of instances, the social circle heuristic made a predictions (i.e., discriminated between the diseases) for only 11.1% of all inferences and was applicable at least once for only 33 participants. Across these 33 participants, the social circle heuristic correctly predicted a median proportion of 79.5% of the inferences (M = 77.0%, SD = 15.9). In comparison, the availability by recall heuristic, which always took account of all instances that the participants reported, predicted a median proportion of 81.8% of inferences correctly (M = 77.6%, SD = 16.6), was applicable for 10.5% of all inferences and made a prediction for 33 participants. Thus, both mechanisms were equally appropriate to predict participants' inferences.

To examine the accuracy of the two mechanisms when applied to the occurrences of the diseases recalled by the participants, I correlated in a first step the total number of

instances that the participants reported in the recall task with the actual number of cases. The correlations were rather high (r = .77, p = .001; r_s = .38, p = .07), indicating that this information was useful for the inference task. In a second step, the predictions of both mechanisms for each participant were compared with the correct choices (i.e., according to the actual incidence rates, averaged values from a 5-year period were used to eliminate year-to-year fluctuations; see Chapter 3.1). The accuracy was defined, separately for each participant, as the number of correct inferences made by the social circle heuristic divided by the number of comparisons where it was applicable. If the social circle heuristic had been strictly applied, it would have reached a median accuracy of 83% (M = 78.0%). In comparison, if availability by recall had been strictly applied, it would have reached a median accuracy of 83% (M = 79.0%). Thus, in line with the computer simulation, both mechanisms had reached a similar accuracy. Moreover, availability by recall retrieved an average of 1.8 (SD = 1.1) instances per choice, whereas the social circle heuristic retrieved only 1.15 (SD = 0.24) instances.[54]

Discussion

In Study 7, I investigated, both in a computer simulation and in an empirical study, a simple inference mechanism that exploits a person's social network as an easily accessible sample space for judging event frequencies in paired comparisons. The results show that the social circle heuristic allows one to judge accurately the environmental frequencies of randomly distributed events. At the same time, this mechanism predicted people's choices rather well compared to a mechanism that relied on the information of individuals' total social network. Thus, the accuracy achieved by the social circle heuristic provides another example for the argument that small samples can be an efficient basis for judgments in the real world (cf. Fiedler & Kareev, 2004; Kareev, 2000; but see Anderson, Doherty, Berg, & Friedrich, 2005; Juslin & Olsson, 2005).

However, the task used in Study 7 had the disadvantage that participants could retrieve only a very small number of instances. As consequence, both the social circle heuristic and availability by recall were applicable for only a small proportion of inferences, and for an even smaller proportion of inferences they made distinct predictions, undermining a rigorous test of the two mechanisms against each other. Therefore the results of Study 7 are inconclusive with regard to the question of whether people actually retrieve only a subsample of the instances from their social network, as hypothesized by the social circle heuristic.

[54] Note that, in contrast to the computer simulations, the comparisons in which the number of recalled instances did not discriminate were not included to determine frugality. But given that both mechanisms had to guess a very similar number of times (both in the simulation and the experiment) this difference in determining frugality should not distort the results.

Overview of Study 8

The major goal of Study 8 was to examine how well the instance-based approach fares compared to the cue-based approach with regard to how well they predict people's frequency judgments. The mechanism competition included the three cue-based mechanisms WADD, EQW and TTB, and the three instance-based mechanisms RECALL, SCH-Group and SCH-Contact described above. For a more rigorous test of the social circle heuristic, a domain was required in which it was reasonable to expect that people would be able to retrieve more instances than for the infectious diseases in Study 7. Sports were chosen for this purpose, and participants had to judge the number of club members that different sports have in Germany. For this domain it was also reasonable to assume that the people could retrieve cues that are positively correlated with the frequency of the event, allowing a test of the two approaches against each other. To be able to test the cue-based mechanisms, I conducted a prestudy asking participants for cues they would use to infer the number of club members for different sports.

Table 3.2.1. *The 25 most popular sports in Germany (in terms of the number of club members: averaged across the years 1997-2001; e.g., Statistisches Bundesamt, 2002) and the number of club members the participants recalled from their social circles.*

Sport	Number of club members (active and passive)	Number of instances recalled by participants
Soccer	6,234,883	180
Gymnastics	4,800,199	12
Tennis	2,085,327	58
Shooting	1,584,931	10
Athletics	851,075	29
Handball	833,345	70
Equestrian	735,229	48
Table tennis	710,267	12
Skiing	677,556	18
Match fishing	650,921	22
Aquatics	633,652	90
Volleyball	530,399	34
Golf	320,630	41
Judo	268,475	49
Bowling	266,538	25
Dance sport	255,190	57
Badminton	230,058	27
Basketball	202,938	93
Sailing	190,577	42
Ice sports	173,625	21
Cycling	153,141	27
Canoe	111,545	8
Karate	106,582	26
Chess	94,172	5
Rowing	78,746	21

Prestudy

Participants. Thirty students from various subject fields (19 females and 11 males, mean age 24 years, 20-37 years) participated. Participation in the prestudy, which was part of an unrelated experiment, was compensated by a payment of €3 (= $3.69 U.S.).

Materials and procedure. Participants were presented with a alphabetically ordered list of the 25 most popular sports in Germany (Table 3.2.1) and asked to imagine that they had to pick out of two sports the one with the higher number of club members. They were asked to write down on a piece of paper those features of sports that would come to their minds and which ones might help to infer which of two sports of the list has a higher number of club members. The task was illustrated by the example of having to infer which of two German cities has more inhabitants. As state capitals are often larger than cities that are not state capitals, the information of whether a city is a state capital would be informative. The participants could mention as many features as they wished. Completing the task took around 15 minutes.

Results

The most frequently mentioned features of sports are reported in Table 3.2.2. These eight features of the sports were used as cues to test the three cue-based mechanisms. Note that whereas for some of the cues a value can be assigned relatively unequivocally to the sports (e.g., whether a sport is a team sport or an individual sport), other cues are more subject to subjective assessment (e.g., "seasonal", "special equipment"). To take such interindividual differences in the assessment into account, in the main study, participants assessed the sports on these cues.

Table 3.2.2. *The cues mentioned by the participants in the prestudy for the task of inferring which of two sports has more club members.*

Cue name	Description
National star	Whether there are famous German athletes for the sport
School	Whether the sport is often performed in sports classes in school
Seasonal	Whether performing the sport is seasonal dependent
Ball sport	Whether the sport is played with a ball or not
Special equipment	Whether special equipment is required to perform the sport
Olympic sport	Whether the sport is an Olympic discipline
Outdoors sport	Whether the sport is mainly played outdoors or indoors
Team sport	Whether the sport is a team sport or an individual sport

Main Study

Method

Participants. Forty students (23 female and 17 male, mean age 24.8, range 20-33) from various subject fields were recruited for the study, which was conducted at the Max Planck Institute for Human Development in Berlin. None of the participants took part in Study 7 or in the pre-study of Study 8. All participants received a part of their payment contingent on their accuracy in the choice task (see below). In addition to an initial fee of €9 (= $11.06 U.S.), they earned 4¢ (=5¢ U.S.) for every correct decision, and 4¢ were subtracted for every incorrect decision.

Materials. The 25 most popular sports in Germany served as the events about which participants had to make frequency judgments. Specifically, the sports had to be compared with regard to the number of club members registered for them. The statistics about the number of club members were obtained from the official statistics (e.g., Statistisches Bundesamt, 2002) and averaged across five consecutive years (1997-2001) to reduce year-to-year fluctuations (see Table 3.2.1).

Altogether the participants completed four tasks: a choice task, and—to be able to determine the predictions of the six candidate mechanisms—a recognition task, a recall task and a cue assessment task. In the two-alternative forced-choice task (*choice task* hereafter), participants were presented with pairs constructed from the 25 sports, altogether 300 pairs. The task was to indicate in each pair for which of the two sports there is a higher number of club members in Germany. In the , participants indicated for each sport whether they had heard of it before. The recognition task was followed by the *recall task*, in which, in a first step, participants reported for each sport and for each of the four circles defined by SCH-Group (self, family, friends and acquaintances) persons who were club members. In a second step, they indicated (on a five-point scale with the categories “Several times a week”, “Once in a week”, “Approximately once a month”, “Around once in six months” and “Less than once in six months”) for each recalled person how often they typically have contact with the person. “Having contact” was defined as talking to the person for at least five minutes, or writing to or receiving a message from the person of around 100 words in length.

In the *cue assessment task*, consisting of four subtasks, the first task was to assess the sports on the eight cues identified in the prestudy. That is, for each of the eight cues, participants assigned to each sport a (binary) cue value (e.g., whether soccer is a team sport or an individual sport; whether badminton is an Olympic sport or not etc). Moreover, participants assessed the predictive direction of the cues, that is, which value of the cues (e.g., team sport or individual sport) indicated a higher number of club members. Based on these assessments the cue values were coded such that a positive value indicated a higher number of club members. Third, participants rank ordered the cues according to their validity. Finally, the cues were presented sequentially in the indicated rank order and participants estimated the validity of each cue. The validity was assessed using a frequency format. Specifically,

participants were instructed to imagine 100 pairs of sports in which one sport had a positive cue value and the other sport a negative cue value. The task was then to indicate in how many out of the 100 pairs the sport with the positive cue value would actually have a higher number of sports members. It was pointed out that "50" meant that the cue was not predictive of the number of club members (i.e., not better than chance).

Procedure. The choice, recognition and cue assessment tasks were presented on a computer, the recall task as a paper and pencil questionnaire. In the choice task the 300 pairs of sports were presented sequentially in blocks of 25 pairs. One sport was presented on the left side of the screen, the other on the right side on the screen. The order in which the sports appeared within a pair was random, as was the order in which pairs were presented (with each participant receiving an individual random order). The task was to choose the sport with the higher number club members in Germany by pressing one of two designated keys on the keyboard. Participants were instructed to keep the index fingers of the right and the left hands positioned on the response keys for the entire duration of a block. The response time in the choice task was recorded. In the recognition task the sports were presented sequentially and the participants indicated whether they had heard of the sport before. The choice task, recognition task, recall task, and cue assessment task were always administered in this order.

Results

I start by reporting participants' accuracy in the choice task. The six mechanisms are then evaluated in two respects. First, I examine how well the information underlying the two approaches (i.e., the number of recalled instances and the cues) reflected the criterion variable in the choice task, that is, the number of club members. This is followed by the analysis of how often the six mechanisms, when they were applicable, made a correct inference (i.e., pointed to the sport with the higher number of club members). I refer to this aspect as the accuracy of the mechanisms. Second, I examine how often the six mechanisms, when they were applicable, predicted participants' choices correctly. I refer to this aspect as the fit of the mechanisms.

Participants' accuracy in the choice task. Participants picked the correct sport in 62.9% (*SD* = 6.3; range 48.7%-77%) of the cases, earning them, on average, €3.10 (= $3.81 U.S.) (*SD* = 1.52) in addition to their initial fee.

Recognition. Shooting ("Schützen") was not recognized by 13 of the 40 participants and match fishing ("Sportfischen") was not recognized by one participant. For comparability, only those comparisons were included in the test of the six mechanisms in which both sports were recognized.[55]

[55] That is, the 2.8% (averaged across participants) of the comparisons where the recognition heuristic was applicable were modelled separately. Across the 14 participants where the recognition heuristic was applicable at least once, on average 59.2% (*SD* = 33.4) of the choices in which it was applicable were in line with it. Although this is a rather low adherence rate, note that the recognition validity α was only .14. In light of this low value of α, it appears that recognition was nevertheless an influential cue in the inferences.

Actual and estimated cue validity. Table 3.2.3 reports the results of the cue assessment task. Concerning the predictive direction of the task, there was a strong consensus among the participants. Only for the cue "outdoors sport", the consensus was less strong, with 24 participants judging outdoor sports to be more popular. To calculate the cue validities (i.e., how well the eight cues allowed to predict the sports with a higher number of club members) two methods were used. First, the validities were determined by using the modal cue value of each sport and the modal direction of each cue. I calculated how often the sport with a positive cue value, when it was paired with a sport that had a negative cue value, had in fact more club members (according to the statistics). Moreover, I calculated for each cue its discrimination rate (DR), which expresses the proportion of times (of the 300 comparisons) the sports had different values on a given cue. The results are shown in Figure 3.2.4 and Table 3.2.3. The cues "national star" and "school" were the most valid ones and the "team sport" cue was least valid. As a second method, since participants differed with regard to both the cue values they assigned to the sports and the cues' directions, I determined the cues' validities separately for each participant. Based on these subjective validities, the "ball sport" cue achieved the highest average validity, followed by the cues "school", "national star", and "seasonal". The "team sport" cue had the lowest average validity.

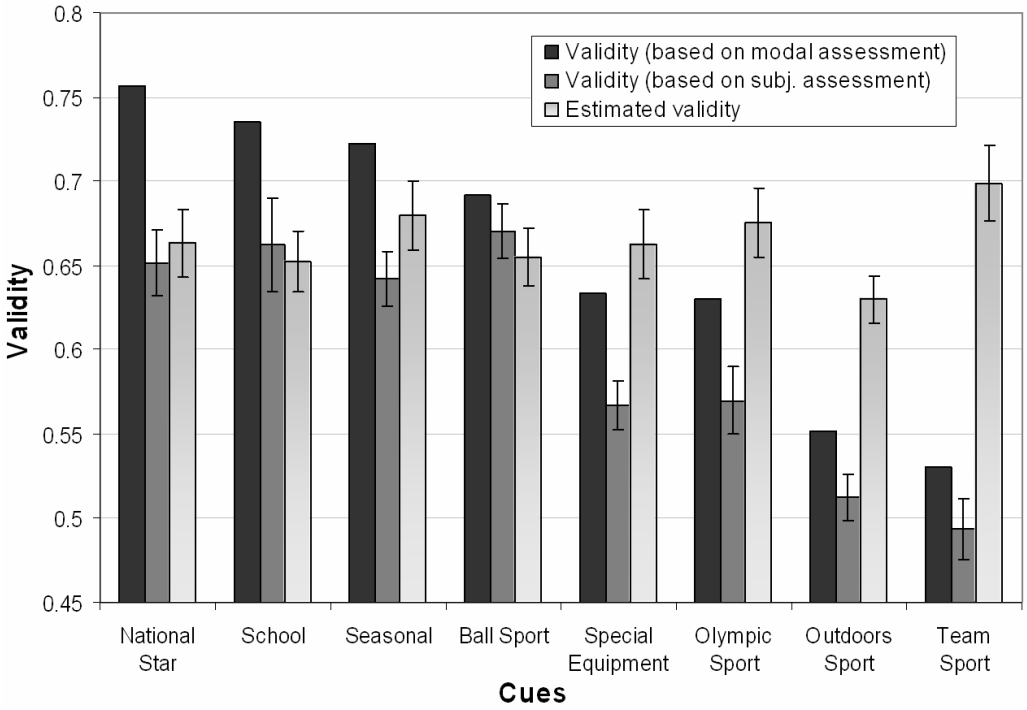

Figure 3.2.4. Actual and estimated validities of the cues identified in the prestudy. The actual validities are depicted both as based on the modal cue values and directions (dark grey bar) and based on participants' individual subjective cue values and direction (medium grey bar). Bars show standard errors.

Did participants' estimated cue validities match these actual subjective validities? Shown in Figure 3.2.4 and Table 3.2.3, there was a substantial discrepancy between the estimated and the actual cue order. Although the "outdoors sport" cue, was correctly estimated to be of low validity (with a median rank of 6 it was assigned the lowest rank and had an actual validity of .55), the "team sport" cue was estimated to be the most valid cue but had the lowest actual validity. Overall, there was no correlation between the cue validities calculated based on the participants' subjective assessments and the estimated validities by the participants, and this was true both on the aggregate level (r = -.17, p = .68; r_s = -.31, p = .46) and on the individual level (average across participants: r = -.03; r_s = .07). When averaged across participants, the mean estimated validity of the cues did not vary much (range 63-69.9), indicating that there was no marked systematic trend in the estimates.

Table 3.2.3. *Direction, validity, and discrimination rate of the cues. The direction refers to the direction that the majority of participants indicated to be more valid. The numbers in the brackets express how many of the 40 participants agreed with the modal direction assessment. The validities based on subjective assessment were calculated using the cue values and the cue direction indicated by the individual participants, as were the discrimination rates. DR discrimination rate.*

	National star	School	Seasonal	Ball sport	Special equipment	Olympic	Outdoors	Team sport
Direction[a]	+ (37)	+ (34)	- (39)	+ (38)	- (37)	+ (32)	+ (24)	+ (34)
Validity[a]	.76	.74	.72	.69	.63	.63	.55	.53
DR	.52	.45	.42	.45	.50	.33	.51	.33
Validity[b]	.65	.66	.64	.67	.57	.57	.51	.49
SD	.12	.18	.10	.10	.09	.13	.09	.11
DR	.48	.45	.40	.41	.49	.43	.50	.40
SD	.04	.05	.10	.07	.04	.08	.03	.08
Estimated validity (*Mdn*)	65	64.5	65	65	65	65	60	72
Rank (*Mdn*)	5	5	4	5	4.5	4	6	3.5

[a] Based on modal subjective assessment.
[b] Based on subjective assessment of the individual participants.

How well did recalled instances reflect the number of club members in Germany? Eight participants indicated that they were members in a sports club (aquatics 3, basketball 2, judo 2, soccer, match fishing, athletics, dancing, each 1; three participants reported being a club member for two sports). The numbers of club members recalled by the participants from their social networks for each sport are reported in Table 3.2.1. Although the number of recalled instances reflected the actual distribution of club members rather well (r = .53, p = .003; one-tailed), it did worse in capturing the differences in ranks among the sports (Spearman r_s = .26, p = .10; Goodman-Kruskal γ = .18, p = .11, both one-tailed[56]). Therefore,

[56] Using the probabilistic interpretation of Goodman-Kruskal's γ, the likelihood that a sport with a higher number of recalled instances has also more club members in the population is $P_\gamma = 0.5 + = 0.5 \times .18 = .59$ (Nelson, 1984; Gonzalez & Nelson, 1996).

one can see from this ecological analysis that though better than chance, the number of recalled instances was not highly predictive of the number of club members.

How accurate were the six mechanisms to infer the higher event with the higher frequency? I examined for each mechanism how often, when they made an unambiguous prediction, they correctly predicted the sport with the higher number of club members. For deriving the predictions of the cue-based mechanisms the subjective cue assessments (in terms of cue values and predictive directions) were used. The results are shown in Table 3.2.4. The two compensatory cue-based mechanisms EQW and WADD achieved the highest accuracy. When they were applicable, they made a correct inference in 64% and 62% of the cases, respectively. All three instance-based mechanisms achieved a worse accuracy than the cue-based mechanisms. The best cue-based mechanism EQW was substantially better than the best instance-based mechanism availability by recall, which achieved an accuracy of 57.6% (p = .001, according to a sign-test). As in Study 7, availability by recall and SCH–Group (57.1%) achieved very similar levels of accuracy, though the latter was considerably more frugal: whereas availability by recall retrieved, averaged across choices and participants, 3.4 instances per choice (SD = 1.8), SCH-Group retrieved only 1.7 instances (SD = 0.6).

Note from Table 3.2.4 that the mechanisms differed considerably with regard to their applicabilities (i.e., how often they made an unambiguous prediction). While the cue-based mechanisms made a prediction for almost all comparisons, the instance-based mechanisms made a prediction for only about half of the comparisons, so that the accuracies of the models are hard to compare. Figure 3.2.5 shows the mechanisms' accuracies when focussing only on the subset of comparisons for which all mechanisms made a prediction (mean applicability = 39%, SD = 18). When comparing the mechanisms' accuracies only for this subset of pair comparisons, the same pattern as reported above emerged.[57]

Table 3.2.4. *Fit (i.e., proportion of correctly predicted choices by the mechanisms when they made a prediction), accuracy (i.e., the proportion of correct inferences), and applicability (i.e., proportion of comparisons an unambiguous prediction was made) of the six mechanisms.*

Mechanism	% of correctly predicted choices (M)	SD	Applicability (M)	SD	Accuracy (M)	SD
Cue-based mechanisms						
WADD	66.6	8.5	.94	.04	.62	.05
EQW	69.0	9.2	.81	.06	.64	.06
TTB	62.7	9.1	.95	.04	.59	.07
Instance-based mechanisms						
RECALL	71.1	10.1	.53	.20	.58	.14
SCH-Group	70.0	10.1	.56	.22	.57	.13
SCH-Contact	69.2	9.6	.57	.22	.57	.13

[57] The accuracies on this restricted set were 0.66, 0.66, 0.63, 0.60, 0.59, and 0.58 for WADD, EQW, TTB, RECALL, SCH-Group, and SCH-Contact, respectively.

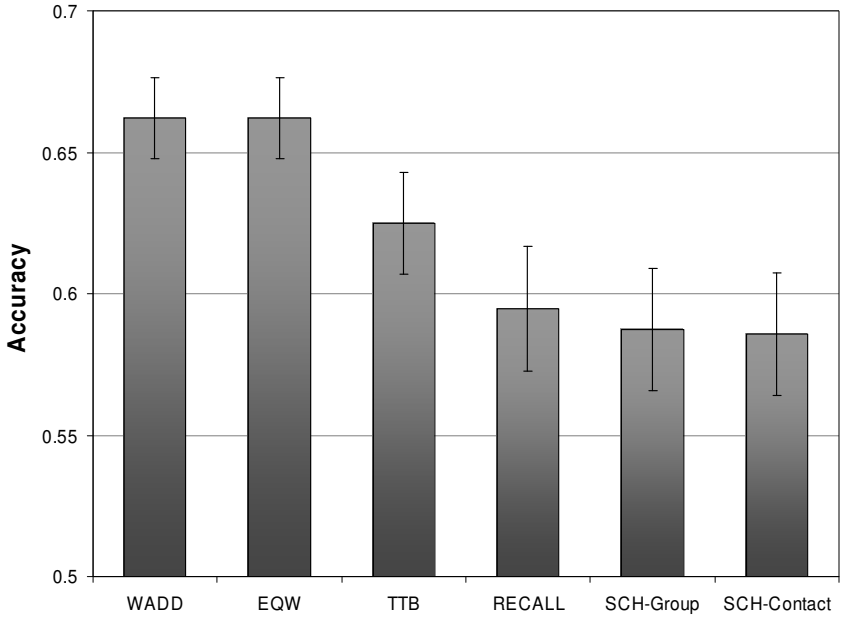

Figure 3.2.5. Accuracies of the six mechanisms on the subset of comparisons where all of them were applicable. Bars show standard errors.

Which mechanism predicted participants' choices best? I now turn to the central question of how well the candidate mechanisms accounted for participants' choices. Could the more accurate cue-based approach also predict participants' choices better, as one would predict when arguing that people's decision processes are adaptive? As described above, for a substantial proportion of inferences the mechanisms did not make an unambiguous prediction. Rather than always having the mechanisms guess in these cases, I analyzed the fit of the mechanisms in two steps. First, I examined the mechanisms' fit irrespective of how often they made an unambiguous prediction. That is, I excluded the cases in which the mechanisms did not make an unambiguous prediction. In a second step (see section Combination of cue-based and instance-based approaches) I included all cases and examined compound mechanisms that either guessed or switched to the alternative knowledge base when the mechanisms did not make an unambiguous prediction.

The average (across participants) proportions of correctly predicted choices (excluding cases where the mechanisms were not applicable) are reported in Table 3.2.4. Two instance-based mechanisms reached the highest fit: availability by recall and SCH-Group correctly predicted 71.1% and 70% of the choices, respectively. Of the cue-based mechanisms, EQW reached the highest fit (69%).[58] However, as reported above, the mechanisms differed in the proportion of choices were they made an unambiguous predictions (see Table 3.2.4). The higher applicability of the cue-based mechanisms could put them at a disadvantage if the

[58] Importantly, the fits of the mechanisms were higher than the participants' accuracies (i.e., their percentage of accurate choices) when the mechanisms were applicable, showing that the mechanisms considered accounted for (some of the) the errors participants made. Specifically, participants' average percentage of correct choices were 63.6%, 64.5%, 63.6%, 64.9%, 65.0%, and 64.9% for the subset of choices of TTB, EQW, WADD, RECALL, SCH-Group, and SCH-Contact, respectively.

comparisons for which they made a prediction beyond the ones for which the instance-based mechanisms made a prediction were harder to predict. To level the playing field, I compared the mechanisms on the subset of comparisons where all mechanisms made an unambiguous prediction. The results are shown in Figure 3.2.6. Although the differences among the mechanisms indeed decreased—suggesting that the mechanisms often made the same predictions—the general picture still held. Availability by recall and SCH-Group predicted the choices best (72.3% and 71.9% correct predictions, respectively). From the cue-based mechanisms, EQW and WADD (which were indistinguishable as they almost always made the same prediction), showed the best fit (both 70.2%), but none of the cue-based mechanisms reached the fit of the instance-based mechanisms (SCH-Contact predicted 70.7% of the choices correctly). TTB achieved the lowest fit and predicted only 65.8% of the choices correctly. The difference between availability by recall and EQW (and WADD) amounted to a small to medium effect size of $w = .17$[59] (Cohen, 1988), though it did not reach conventional levels of significance (sign test $p = .19$).

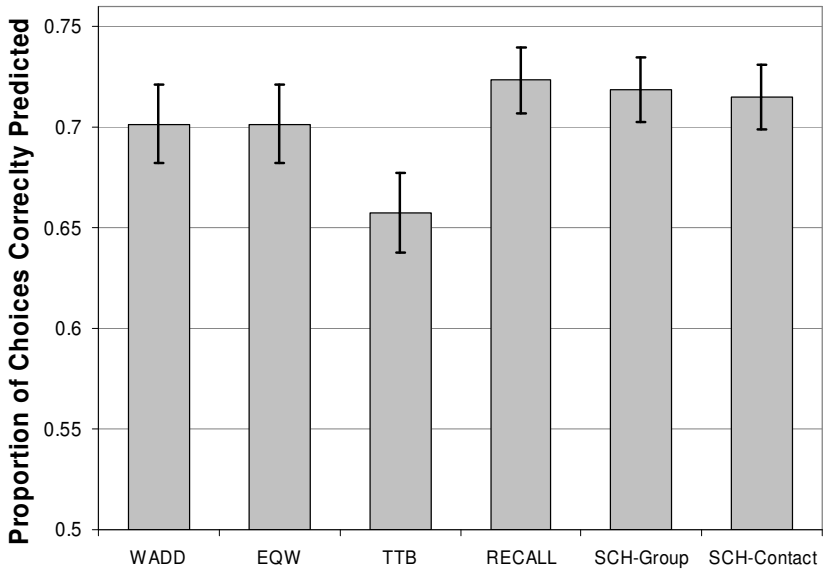

Figure 3.2.6. Fit of the six mechanisms in predicting participants' choices, considering only the cases where the mechanisms were applicable. Bars show standard errors.

From the analysis on the aggregate level, one can thus conclude that the instance-based mechanism availability by recall was best in predicting participants' choices, but beat its competitors by only a small margin. To corroborate these results, trace possible individual differences in strategy use, and tackle the problem that the mechanisms often made the same predictions, the participants were additionally classified individually to the different

[59] There were 23 positive and 14 negative differences, and 3 ties. Eliminating one tie and splitting the other 2 across the two mechanisms (cf. Bortz, Lienert, & Boehnke, 2000) yielded a proportion for availability by recall of .62, which yielded an effect size—relative to chance = 0.5—of $w = \sqrt{\frac{(0.5-0.62)^2}{0.5}} = .17$.

mechanisms. For each participant the six mechanisms were compared in pair-wise contests (6 × 5 / 2 = 15 contests). These contests were based only on those comparisons that allowed to distinguish between a given pair of mechanisms, that is, those comparisons for which both mechanisms were applicable and for they made different predictions (see Chapter 3.1). A participant was assigned to the mechanism that "won" the highest number of contests (maximum of 5). After this classification, each mechanism received one point for every participant assigned to it. If a participant could not unequivocally be assigned to one mechanism (e.g., because two mechanisms both won the same number of comparisons), the point was equally distributed across the tied mechanisms (i.e., 0.5 if two mechanisms were tied, 0.33 if three mechanisms were tied etc). Figure 3.2.7 shows the summed points for each mechanism. Convergent with the aggregate-level analysis, availability by recall and SCH-Group emerged as the winners and received the highest number of points (12.5 and 10.5, respectively, a non-significant difference: $\chi^2 = .17$, $p = .84$ [exact significance]). From the cue-based mechanisms, EQW was best (5.33) and received a slightly higher number of points than WADD.[60] Although this individual classification should be seen as just a crude approximation, it generally converges with the results on the aggregate level.

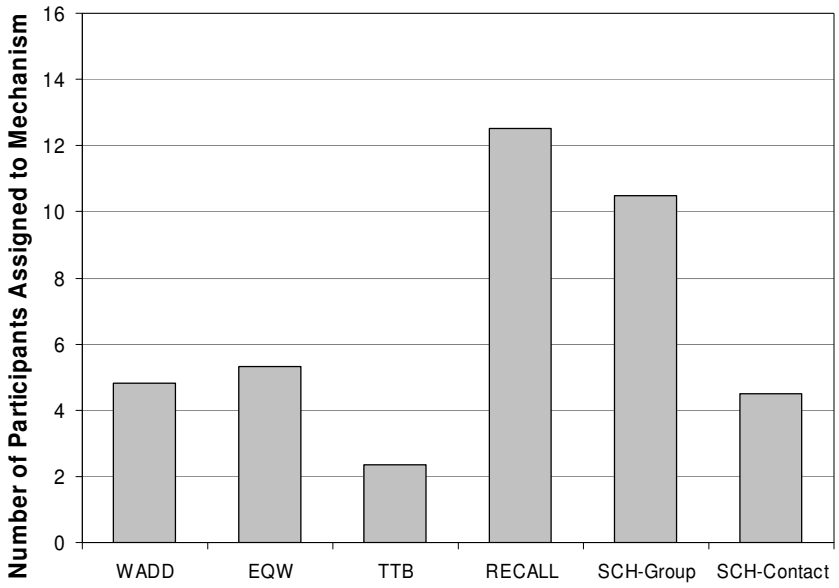

Figure 3.2.7. Classification of the 40 participants to the six mechanisms (see text for details).

To summarize, both on an aggregate and on an individual level of analysis the instance-based mechanisms consistently achieved a better fit than the cue-based mechanisms. Of the instance-based mechanisms availability by recall and SCH-Group were best able to predict participants' choices, whereas SCH-Contact always achieves the lowest fit. Of the

[60] EQW outperformed WADD only indirectly, however. Both mechanisms were indistinguishable in a direct comparison and the higher number of points EQW was only due to its having a better fit than WADD when compared to other mechanisms.

cue-based mechanisms, EQW and WADD achieved the best fit, but were hardly distinguishable, as they very often made the same prediction.[61] Finally, TTB clearly fell behind and did not seem to be able to capture participants' choices. Altogether, the results suggest that the instances people can retrieve from episodic knowledge are used as a primary knowledge base for inferences about event frequencies in the population, whereas predictive information from semantic knowledge is used less.

Examining the inference process: response times. The analysis so far has examined the candidate mechanisms on the outcome level, that is, with regard to whether the predicted choices coincided with the actual choices of the participants. But some of the mechanisms make different predictions also on the process level, and response times can be used to test these predictions. In particular, the social circle heuristic assumes that instances are retrieved sequentially by circle. As a consequence, the response time of choices following the social circle heuristic should differ as a function of the number of circles from which instances were retrieved. Because SCH-Contact clearly failed to account for the choices, I focus on the circle definition of SCH-Group exclusively. If the assignment to the six different mechanisms also captures differences in terms of the processes, then the response times of the participants assigned to SCH-Group should differ as a function of the circle that is predicted to determine the choice, whereas for participants assigned to the instance-based mechanisms there should be no such difference. It is unclear whether there should be a difference for availability by recall. Participants assigned to availability by recall do not seem to generally follow the stopping rule hypothesized by the social circle heuristic (i.e., stop retrieving instances as soon as the number of instances allows to discriminate between the events). The search rule (i.e., retrieval by circle), however, might still be valid for them as well. Thus, the participants assigned to availability by recall may not generally, but occasionally, stop retrieving instances prematurely (recall that no participant followed availability by recall in a deterministic fashion), and then the response times of choices made by these participants could still differ among the circles.

For the following analyses, the response times in the choice task were natural log-transformed and z-standardized for each participant (to reduce inter-individual differences). Also, the choices, rather than the participants were chosen as unit of analysis. Only comparisons were included that involved two recognized sports and those for which the number of instances differed.

I analyzed the response times as a function of the circles which SCH-Group predicted to determine the choices, separately for participants assigned to SCH-Group, RECALL, and the cue-based mechanisms (which were collapsed to enhance power). The number of instances summed for both events, the difference between the number of instances (both natural log-transformed to reduce the skewed distribution), and the difference between the number of positive cue values were included as covariates. In line with the prediction derived

[61] In fact, in the direct contest, the two mechanisms made a different prediction only once with two participants. One was in favor of EQW, one in favour of WADD.

from the individual classification, there was a main effect for circle for the participants assigned to SCH-Group (F[3, 1113] = 6.16, p = .001), but no main effect for circle for the participants assigned to the cue-based mechanisms (F[3, 2241] = .81, p = .49). Also, as can be seen from Figure 8, the response times followed roughly the monotonic trend predicted by the social circle heuristic. Suggesting that the circles also had an effect with the participants assigned to availability by recall, there was a main effect for circles for these participants as well (F[3, 1162] = 3.86, p = .009).

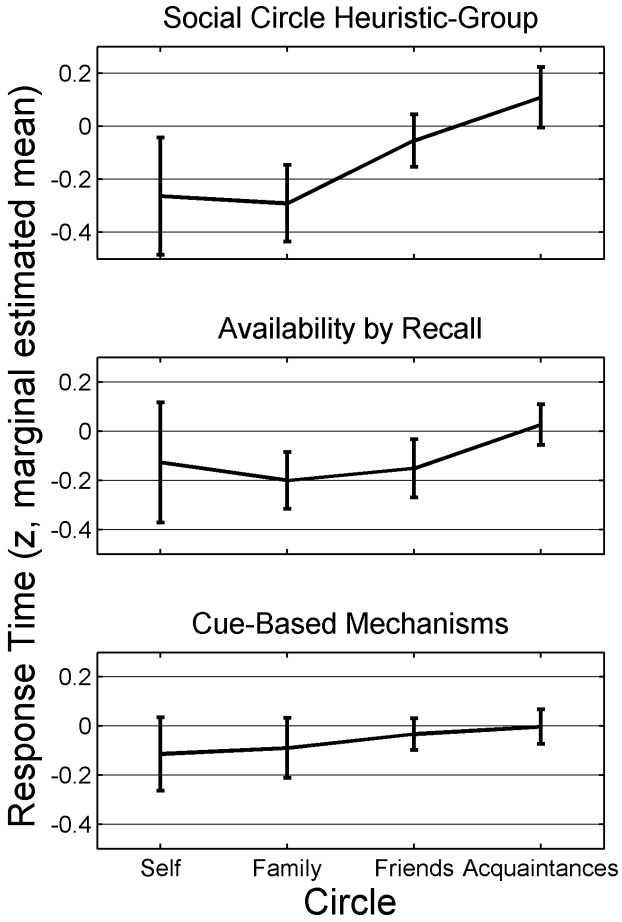

Figure 3.2.8. Response times as a function of the circles predicted to determine the choice, separately for participants assigned to SCH-Group, availability by recall, and the cue-based mechanisms. The bars indicate standard errors.

Combination of cue-based and instance-based approaches. All previous analyses excluded the cases in which the mechanisms did not make an unambiguous prediction. This was particularly often the case for the instance-based mechanisms, and it might be interesting to know how participants proceeded in this situation. Was a decision then made by guessing? Alternatively, instead of making a random choice, people might switch to the alternative knowledge base to make an inference. To examine this assumption, I focus on the mechanisms of each approach that was best in predicting the inferences, which are EQW and

availability by recall. First, I consider the fit of the mechanisms assuming that cases in which they do not make an unambiguous prediction are solved by picking one of the two sports randomly. Second, I determined the fit of two compound mechanisms. First, *RECALL-EQW* predicts that an inference is made following availability by recall, and in case the mechanism does not allow for an unambiguous prediction, an inference is made based on EQW. In case an unambiguous prediction is still not possible, a random choice is made. Second, *EQW-RECALL* predicts that an inference is made following EQW, and in case EQW does not allow for an unambiguous prediction, an inference is made based on availability by recall. As above, in case an unambiguous prediction is still not possible, a random choice is made.[62]

Which mechanism matched participants' choices best, now looking at all choices? A strategy classification method proposed by Bröder and Schiffer (2003b) was used to assign participants to the four mechanisms (see Appendix B for a more detailed description of the classification method).[63] The result of the classification is reported in Table A in Appendix B (rather than the likelihoods, the fit measure G^2 is reported, with lower values indicating a better fit). Twenty of the 40 participants were classified as using RECALL-EQW, 9 participants were assigned to EQW (guess), 5 to EQW-RECALL and 5 to RECALL (guess).[64]

Since the largest number of participants were assigned to RECALL-EQW it can be inferred that most participants first accessed episodic knowledge and based an inference on the number of recalled instances in their social networks. Cues, however, were also accessed and an inference based on them when the number of instances did not discriminate between the sports (and a sport was picked randomly when even cues did not discriminate).

Finally, note that one direct process implication of the strong support that RECALL-EQW obtained is that the response time of cases where the sports could be discriminated only after consideration of the cues should be longer than cases where the sports could be discriminated based on the number of retrievable instances. These differences in response time were indeed found. Choices were made significantly faster when the number of instances allowed to discriminate between the sports than when the sports could only be discriminated after consideration of the cues (using the difference of sum of positive evidence as covariate; estimated marginal *M*s = -0.07 vs. 0.06; $F(1, 10710) = 44.8$, $p = .001$).

---

[62] Concerning the accuracy of the different mechanisms, availability by recall and EQW (with random choice in unambiguous cases) made 54% and 60.8% correct inferences, respectively, and RECALL-EQW and EQW-RECALL made 58% and 60.9% correct inferences, respectively. EQW-RECALL and EQW (followed by guessing when no unambiguous prediction could be made) were the mechanisms leading to the highest accuracy.

[63] This method was not possible for the analysis above, which tested the mechanisms only when they discriminated, as with this method, all mechanisms have to be tested on the same set of comparisons.

[64] Bröder & Schiffer, 2003b, proposed to classify participants as "guessing" if the random error ε_k of the best-fitting mechanism exceeds a certain high value, for instance, $\varepsilon_k = .40$. One participant—# 38 with $\varepsilon_k = .51$ [see Table A] exceeded this threshold and was classified to a pure guessing strategy.

General Discussion

The main goal of this chapter was to test two approaches—the cue-based approach and the instance-based approach—for predicting people inferences about event frequencies. Each approach was represented by three different mechanisms. The mechanisms differed, apart from the type of information used, in the amount of information considered and the way information was searched for. In Study 7, I first focused on the instance-based approach, and formulated a new instance-based mechanism—the social circle heuristic—that is based on a sequential, but often incomplete retrieval of instances in a person's social network. First, by means of a computer simulation I demonstrated the competitive accuracy of the social circle heuristic compared to availability by recall, which retrieves all available instances. Second, the social circle heuristics did equally well as the availability by recall mechanism in predicting participants' inferences in the conducted experiment. In Study 8, the cue-based approach and the instance-based approach underwent a rigorous comparison test, to examine which predicted participants' inferences about event frequencies best.

The results of this direct comparison showed both on an aggregate and on an individual level that the instance-based approach was most successful to predict participants' inferences. Specifically, availability by recall, an instance-based mechanism that makes an inference based on all instances a person can recall from her social network, achieved the best fit. Statistically not reliably different from this mechanism, a version of the social circle heuristic in which the circles are defined by the social group to which a social network member belongs achieved the second best fit. The individual analysis suggests this mechanism predicted the choices of a substantial proportion of participants best. That is, there is evidence that some people make an inference based on only a subset of all instances. Moreover, a process analysis based on response times suggests that social circles guide the retrieval of instances, even if the search process is not generally stopped as soon as the number of instances in a circle allows to distinguish between the events.

None of the cue-based mechanisms was able to compete with the two best instance-based mechanisms. Of the three cue-based mechanisms, the equal weight linear mechanism showed the best fit (although it was practically indistinguishable from a weighted linear model), whereas the simpler Take The Best heuristic failed clearly to predict participants' choices. The notion that in the information search process people start be considering instances in their social networks was further corroborated in a test in which the cue-based and instance-based approaches were combined. This analysis suggested though not primary, cues come into play when the number of retrievable instances does not allow for an unambiguous inference. The process-oriented analysis corroborated these results. Importantly, however, although the cue-based mechanisms did not predict participants' choices well, they were the more accurate approach to make inferences about the number of sport club members, that is, they were better at predicting the sport with the higher number of club members. In the following, I discuss the implications of the findings.

The Primacy of Instances as the Basis for Frequency Judgments

Why was the instance-based approach more successful in predicting participants' inferences about event frequencies? Its better fit is puzzling in light of the higher accuracy of the cue-based mechanisms.[65] Assuming that people were sensitive to the differences in accuracy, an effort-accuracy perspective (e.g., Beach & Mitchell, 1978; Payne et al., 1993) would suggest that the cognitive costs associated with the cue-based mechanisms must have outweighed the higher accuracy they afforded. Although it remains unclear whether participants knew that the cue-based mechanisms were more accurate, let me I offer some speculations as to possible factors producing such higher costs.

First, due to its more abstract nature, the semantic knowledge that cues represent might be less readily activated than instances of the critical events. Also, the computations underlying cue-based inferences involve multiple steps (evaluation of cues in terms of the predictive direction, assessment of cue validity, integration of cue values and validity etc). Instances of an event, by contrast, are probably directly activated by the name of the event, come in a common currency, and could thus be easier to combine for an inference. Finally, instances are typically learned sequentially. As a consequence, the frequency information coded by them is represented as natural frequencies (Gigerenzer & Hoffrage, 1995), which can foster probabilistic reasoning (Hoffrage et al., 2000). Together, these factors could lead to lower cognitive costs and thus a primacy of instances as the basis for judging how often events occur in a population.

Ecological Rationality and the Adaptive Use of Cues

A central thesis in Brunswikian approaches to judgment and decision making is that people are well adapted to the validities of proximal information to predict a distal criterion (Brunswik, 1943). For instance, it is central in Hammond's Social Judgment Theory (Hammond et al., 1975), PMM theory (Gigerenzer et al., 1991) and the fast and frugal heuristics approach (Gigerenzer et al., 1999). Two of the results that were obtained, however, challenge the assumption that people are fine tuned to the predictive value of information in their environment. The first is the discrepancy between accuracy and use of the cue-based approach discussed above, which, however, could be accounted for by higher cognitive costs associated with the cue-based approach. The second result troublesome for the notion of adaptive decision making is that participants' judged cue orders (in terms of validity) were unrelated the actual cue orders. As the cues did not refer to obscure features of the sports it is not very plausible that the discrepancy between estimated and actual cue order was due to

[65] Which is surprising given that participants' estimates of the cue validities were relatively inaccurate; a more veridical perception of the cue validities might have led to a higher accuracy of the cue-based approach. But note also that participants' insensitivity to the actual cue validities did not harm to the equal weight or the weighted additive mechanism, as in the former only the direction of the cue validities are taken into account and for the latter only large differences in cue validity lead to different predictions. Take The Best, however, is highly sensitive to inaccurate cue validities, which might explain why it achieved the lowest accuracy of the three cue-based mechanisms.

participants' lack of expertise. Moreover, note that as the assessments were given after the choice task, participants had some opportunity to consider the validity of the cues. Although there were no systematic trends in the estimates, the fact that the "team sport" cue was estimated to have the highest validity might suggest that estimates were based on some kind of plausibility analysis.

Overall, the results obtained indicate that at least some intentional learning (with feedback) might be required to learn the correct cue order. Overall, the apparent difficulties resonate with findings are in line with work in the multiple cue probability learning (MCPL) literature that has shown that people have trouble learning cue validities correctly even when provided with ample opportunity to do so (e.g., Connolly & Gilani, 1982; Connolly & Serre, 1984; Goldberg, 1968). Note, moreover, that the observed inaccuracies make it difficult for mechanisms (such as TTB) that hinge on the processing of cues in an approximately accurate cue order to achieve accurate inferences.

The Recall Principle: Old Wine in a New Bottle?

As pointed out in the introduction, the notion that when making frequency judgments, people try to retrieve instances—the recall principle—is one of the mechanisms described as the availability heuristic (Tversky & Kahneman, 1973). One should emphasize, however, that the recall principle is not equivalent to the availability heuristic. First, as mentioned above, whereas the availability heuristic as originally proposed is consistent with two distinct processes—namely ease of recall and actual recall (see Betsch & Pohl, 2002)—the recall principle refers exclusively to recall. Second, by specifically focussing on recall, the recall principle operates on a knowledge base that is less prone to bias than ease of retrieval. Distortions produced by relying on frequencies in the sample of people one knows to estimate frequencies in the population are primarily produced by sampling error and therefore stems rather from the external environment than from a biased cognitive process (Fiedler, 1996, 2000). The recall principle is much less susceptible to distortions due to memory-related factors and distortions caused through selective news coverage, from which the availability heuristic has been claimed to suffer. For instance, the availability heuristic would extend search to the virtual circle, which includes instances reported by the mass media (cf. Chapter 3.1).

Relation of the Recall Principle to Minerva-Decision Making

An important development concerning the possible memory processes underlying frequency judgments was Dougherty et al.'s (1999) simulation of "availability effects" (among others) in Minerva-Decision Making (MDM), a multiple-trace memory model. It is therefore important to explicate the relation between the instance-based approach and MDM.

First, it should be noted that both the instance-based and cue-based approaches, and MDM describe decision making as a function of memory. One basic difference, however, is that whereas MDM assumes that inferences are based on some general process (i.e., echo

intensity) arising from activation of memory traces, both the instance-based and cue-based approaches describe judgments based on higher level cognitive algorithms (cf. Dougherty et al., 1999). Conversely, only MDM specifies possible memory retrieval processes. An interesting dimension of comparison is the issue of information loss, a central issue in MDM (Dougherty, 2001). In MDM (as in Fiedler's, 1996, BIAS model) information loss is due to the fallibility and imprecision of memory, leading to the probabilistic, rather than deterministic, relationship between mind and environment. Information loss is considered in the cue-based and instance-based approaches, too, but here information loss and the ensuing uncertainties stem from a probabilistic relationship that is already in the environment, namely between the proximal information (cues or instances) and the criterion. Even without information loss in the cognitive system there would be uncertainty left.

In spite of these differences, if traces are taken to represent instances of the events (rather than repeated occurrences with the same person; e.g., Fiedler, 1996), it is possible to instantiate the instance-based approach in MDM. The traces contain the knowledge one has about the individuals in one's social network. Availability by recall could be simulated by weighing all traces (or instances) equally, for instance, by only considering the component of the trace vector that represents the type of event (e.g., the type of sport). Moreover, events that are not exactly of the same type would have to be ignored completely (e.g., by a very steep function linking similarity to activation, leading, for instance, to the ignorance of the similarity between basketball and handball). The social circle heuristic could be simulated in MDM by coding the location of a trace (i.e., in which social circle) such that the weight of an instance decreases in a noncompensatory way from central to the peripheral circles. In spite of these similarities, MDM does not assume sequential search of instances and therefore cannot explain differences in the response time.

Conclusion

For many frequency judgments outside the laboratory we have no experience of all occurrences to which the judgment refers, forcing us to make an inference based on uncertain information. In applying various mechanisms that might apply in this situation I have brought together concepts developed in the tradition of the availability heuristic and concepts developed by the Brunswikian tradition, which are typically studied in isolation from each other. The paradox was observed that although people can come up with information that would help them to correctly predict an important aspect of the environment, namely event frequencies, this information seems to be considered only secondarily. Moreover, I highlighted the interplay between instance-based and cue-based inference mechanisms for frequency judgments, thus demonstrating the usefulness of considering both approaches in combination. With this integration, I hope to have contributed to a further understanding of the processes underlying people's use of the small sample of instances when judging the frequency of occurrence of events.