Journal of Experimental Psychology:
Learning, Memory, and Cognition
2005, Vol. 31, No. 4, 621–642

Copyright 2005 by the American Psychological Association
0278-7393/05/$12.00 DOI: 10.1037/0278-7393.31.4.621

Judgments of Risk Frequencies: Tests of Possible Cognitive Mechanisms

Ralph Hertwig
University of Basel

Thorsten Pachur and Stephanie Kurzenhäuser
Max Planck Institute for Human Development

How do people judge which of 2 risks claims more lives per year? The authors specified 4 candidate mechanisms and tested them against people's judgments in 3 risk environments. Two mechanisms, *availability by recall* and *regressed frequency,* conformed best to people's choices. The same mechanisms also accounted well for the mapping accuracy of estimates of absolute risk frequencies. Their nearly indistinguishable level of performance is remarkable given their different assumptions about the underlying cognitive processes and the fact that they give rise to different expectations regarding the accuracy of people's inferences. The authors discuss this seeming paradox, the lack of impact of financial incentives on judgmental accuracy, and the dominant interpretation of inaccurate inferences in terms of biased information processing.

Key words: availability, heuristics, health risks, regression toward the mean, risk perception

How the public perceives health risks has been a long-standing concern in the medical community. Writing in 1882 about infectious diseases, William Simpson, the medical officer of health for Aberdeen, Scotland, identified a spurious link in the public mind between high frequency and low lethality:

> It comes out, as a peculiar fact, that the most dreaded diseases are the least fatal, and the least dreaded diseases are the most fatal. . . . Measles, whooping cough and scarlet fever are the most serious, although it is usually considered they do little harm. . . . Their very frequency makes them less dreaded. . . . The disease that comes unexpectedly, and passes over quickly, is looked upon with greater feelings of terror than the disease which may be more fatal, but more common. (cited in Pennington, 2004, p. 28)

As for why the public may erroneously view a disease as harmless, the medical community has been of many (not necessarily contradictory) minds. Whereas Simpson seemed to imply that the ubiquity of some killers caused them to be regarded as a mere nuisance, more than a century later, medical doctor Ronald J. Glasser (2004) suggested that it is their relatively low prevalences that make some public health threats appear innocuous. According to Glasser, the public's current sense of safety is particularly unwarranted because medical historians consider the past few decades an age of "emerging plagues," in which factors such as overpopulation, poverty, global climate change, chemical pollution, and industrial agriculture "conspire to create the conditions for unprecedented death by infectious disease" (Glasser, 2004, p. 36). In support of this conclusion, Glasser painted a grim portrait of the immediate threat in today's America:

> In 1995, 1.7 million American patients contracted hospital-spread infections; 88,000 of these patients died; 70% of the infections were drug-resistant. Each year an estimated 76 million Americans fall ill to food-borne illnesses resulting in approximately 325,000 hospitalizations and 5,000 deaths. Influenza infects 10% to 20% of the U.S. population every year and kills 36,000. (Glasser, 2004, p. 36)

In light of these figures and the potential risk of even worse epidemics in the future, 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 article. Specifically, we focus 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. We pitted these views against each other to assess their relative merits. We began by formulating two cognitive mechanisms implied by each view and deriving specific predictions from each mechanism. The predictions were tested at the level of aggregate frequency judgments and estimates (Studies 1 and 3) and at the level of individual frequency judgments (Study 2).

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

Ralph Hertwig, Department of Psychology, University of Basel, Basel, Switzerland; Thorsten Pachur and Stephanie Kurzenhäuser, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany.

Stephanie Kurzenhäuser is now at the Department of Psychology, University of Basel, Basel, Switzerland.

Our thanks go to Michael Dougherty, Ido Erev, Barbara Fasolo, Gerd Gigerenzer, Ulrich Hoffrage, Konstantinos Katsikopoulos, Derek Koehler, Jörg Rieskamp, Lael Schooler, Peter Sedlmeier, and Thomas Wallsten for many constructive comments. We also thank Anita Todd and Valerie Chase for editing the manuscript.

Correspondence concerning this article should be addressed to Ralph Hertwig, University of Basel, Department of Psychology, Missionsstrasse 60/62, Basel 4055, Switzerland. E-mail: ralph.hertwig@unibas.ch

posed by artifacts such as guns, cars, and electricity 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 directly encoded and that the automaticity of the encoding 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, we 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.[1] 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 [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 about 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. We propose the following definitions of these mechanisms.

Availability-by-recall mechanism. In the context of risk frequency judgments, we 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 we 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

[1] 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.

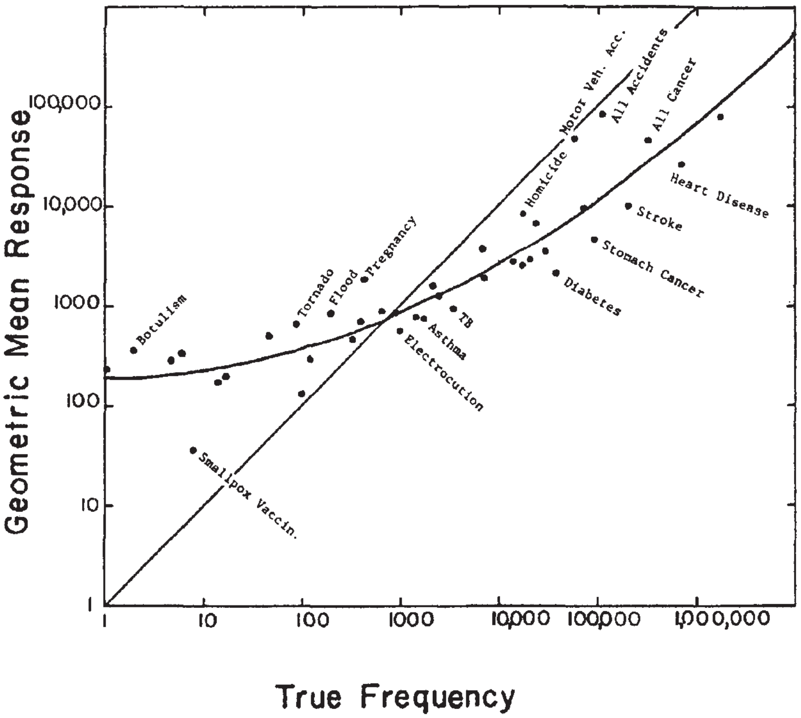

Figure 1. The primary bias, illustrated by the relationship between estimated and actual number of deaths per year for 41 causes of death in Lichtenstein, Slovic, Fischhoff, Layman, and Combs (1978). Each point is the mean estimate (geometric mean) of 39 students. The observation that, for rare causes 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. From "Judged Frequency of Lethal Events," by S. Lichtenstein, P. Slovic, B. Fischhoff, M. Layman, and B. Combs, 1978, *Journal of Experimental Psychology: Human Learning and Memory, 4,* p. 565. Copyright 1978 by the American Psychological Association.

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).[2]

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 (1973) 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 our second interpretation of availability (henceforth referred to as the *fluency mechanism*), we 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. We 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 our view, one elegant proxy is environmental statistics—that is, using tallies of the frequencies with

[2] 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.

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, we 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. We return to this issue later in the General Discussion.

Predictions

In what follows, we 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, we 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 1 lists them and their respective mortality rates in Germany. We 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. We refer to these sets as the *cancer set* and the *infection set*. Table 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.[3] 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 ours; in contrast, in the assorted set, we adopted the composition chosen by Lichtenstein et al. (1978, p. 554). In

[3] 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 1

Entries in the Assorted Set, the Cancer Set, and the Infection Set; Average Annual Mortality Rates (Assorted Set) and Incidence Rates (Cancer and Infection Sets) Averaged for the Years 1996–2000; and Median Estimated Frequencies (Study 3).

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

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 for the years 1997 to 2000 (Arbeitsgemeinschaft Bevölkerungsbezogener Krebsregister in Deutschland, 1999; Robert Koch Institute, 2001). The infection set encompassed 24 infections (dangerous infectious diseases) that by law are notifiable in Germany (*Bundesseuchengesetz*—a law that has recently been revised and now encompasses, for instance, HIV). Note that we dropped rabies from the infection set because there was no single incident during the specified time period.

[a] Included in Study 1.

[b] Not included in analysis.

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, we conducted a pilot study 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.

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:

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

where Choice proportion$_{\text{Risk } a}$ is the proportion of participants who select Risk *a* to be more likely than Risk *b,* and ΣRecalled instances$_{\text{Risk } a}$ and ΣRecalled instances$_{\text{Risk } b}$ are the sum of instances (recalled by all participants) of Risk *a* and Risk *b,* respectively. Here and throughout this article, Risk *a* denotes the event that is, in reality, the more frequent one in a pair comparison. We did not simply assume 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, we employed a probabilistic choice rule to derive choice proportions. That is, we 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*).[4]

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, we 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, we counted the frequency of occurrence with which, for instance, the words *died from breast cancer* were mentioned.[5] They occurred 3,302 times. We 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, we used only the names of the diseases (excluding *died from*). We 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:

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

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

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). We assumed the amount of regression to be 10%. To arrive at this estimate, we 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. We used the results from the latter group to estimate the amount of regression because our focus was on lay judgments. The median amount of regression observed in students' estimates was 10.2%.[6]

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

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

[4] The numerical predictions of the candidate mechanisms for all pair comparisons can be obtained directly from us.

[5] 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/). Our analysis was based on a total of 1,211,000,000 words.

[6] To determine the amount of regression, we 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, we 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, we determined the median amount of regression.

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.[7]

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.[8] 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 cancer or firearm accidents) but also summation categories such as all disease, all cancer, all accidents, suicide, and homicide. For these summation categories, we 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 Regressed average frequency$_{\text{Category } A}$ and 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 we turn to Study 1, 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, we 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 1: Which Mechanism Accounts Best for Judgments of Risk Frequencies?

Study 1 pursued two goals. First, we hoped to replicate the results reported by Lichtenstein et al. (1978). On the basis of this replication, we then would examine which of the candidate processes, if any, could predict people's risk frequency judgments in our study and, by extension, in theirs. Second, we 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 ([$1.26 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—an issue to which we return in the final discussion.

Materials. Table 1 lists the risks included in the assorted set, the cancer set, and the infection set. For all three sets, we 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). We examined the same pairs. From the cancer set, we randomly drew 10 types of cancer and constructed a set of all possible pairs (45). We 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, we included a glossary for some events. If possible, we replaced medical jargon (in the infection and cancer sets) with more commonly used terms. We

[7] The value of, say, breast cancer was calculated as follows: 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).

[8] 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.

Table 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 1			Study 2	
Percentage correct	Assorted set (n = 45)	Cancer set (n = 35)	Infection set (n = 30)	Cancer set (n = 40)	Infection set (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	4.7	8.6	8.1	5.1	5.7
Item difficulty (*Mdn* ratio)	10.9	3.5	72.4	3.2	37.4

consulted a physician to assure the equivalence of medical and colloquial terms.[9]

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

> We ask you 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 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 3, 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 1), for this set we did not force participants 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, we 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 we turn to the test of the mechanisms, we describe the obtained choices in more detail. Table 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.

Why did mean accuracy vary so markedly across sets? We 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, so we 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 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 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.

Which mechanism predicted choices best? To answer this question, we 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 shows the RMSD for each mechanism.[10] 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 2), this prediction is wrong.

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, we computed Spearman rank correlations between predicted and actual choice proportions. As Table 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

[9] In one instance, however, our choices of words went astray. We used the term *food poisoning* (*Lebensmittelvergiftung*) 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. We decided to exclude this item from all analyses, thus reducing the number of pairs in the infection set to 36.

[10] Across all four mechanisms, we excluded 3 pairs the assorted set because their mortality rates turned out to be exactly equally frequent. In addition, for the fluency mechanism, we excluded 17 pairs 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).

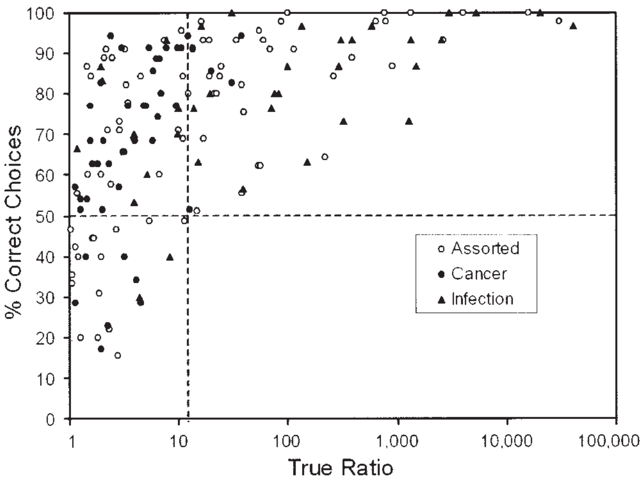

Figure 2. Choice proportion and item difficulty: percentage of participants who correctly identified the more frequent of two risks as a function of the ratio of more frequent to less frequent risk in the assorted set (empty circles), the cancer set (filled circles), and the infection set (triangles). We excluded 28 of the 106 pair comparisons from the assorted set because the actual mortality rate of at least one event was zero.

than the other two mechanisms, except in the cancer set, in which the fluency mechanism performed best.

In sum, we 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 2: A Second Test Involving Individual Responses and Another Definition of Fluency

The poor performance of the fluency mechanism in Study 1 came as a surprise to us. In line with the common wisdom that media coverage shapes people's risk perception (see also Combs & Slovic, 1979), we 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 may be reasons for the inferior performance of the fluency mechanism.

Study 2 was designed to examine the robustness of results of Study 1 by examining an alternative definition of fluency. Specifically, we 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, in press). 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, we 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, we 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, we set the just-noticeable difference to 100 ms (see also Schooler & Hertwig, in press). We do not claim, 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, we also derived individual-specific predictions for the other mechanisms: In the case of the availability-by-recall mechanism, we assumed that if a person recalls more instances of one of two diseases among his or her social circle, then he or she can infer that this disease also has a higher population incidence rate. For the regressed-frequency mechanism, we 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 mecha-

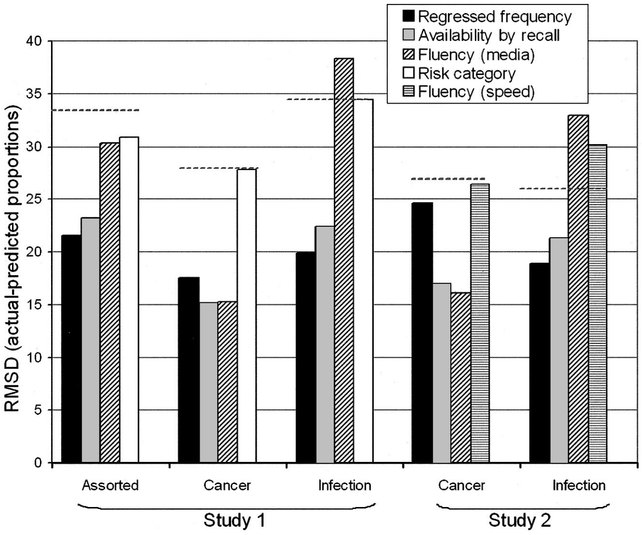

Figure 3. Which mechanism predicted choices best? Root-mean-square deviations (RMSDs) between predictions derived from the four mechanisms and actual choice proportions in the assorted set, the cancer set, and the infection set of Study 1 and the cancer set and the infection set of Study 2. The dotted 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.

Table 3
Spearman Rank Correlation Coefficients Between Actual and Predicted Choice Proportions

	Study 1			Study 2	
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. There are no fluency (speed) values for Study 1 because it was not tested in that study. Except for the negatives, all correlations are statistically significant ($p < .05$, two-tailed).

nism is handicapped as it predicts the same choice across all participants for any given pair of diseases.

Study 2 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 2, we used all 24 elements in the cancer set and the infection set (see Table 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 2). Would the results obtained in Study 1 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 1 (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. We did not include the assorted set 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, we 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 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 we collected the recognition judgments prior to the choices because we were concerned that the reverse order might conflate 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. We deemed this possibility, however, less problematic because it would work in favor of the fluency mechanism, and Study 2's goal was to give the fluency mechanism a second chance. Finally, as in Study 1, 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 we turn to the test of the mechanisms, we 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 2 for more detailed information). The level of accuracy in the infection set was lower than that achieved in Study 1 (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 2 than in Study 1 (37.4 vs. 72.4; see Table 2).

Which mechanism predicted individual choices best? Figure 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, we 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, we 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.

Next, we 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 5, top panel), 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 5, bottom panel), in contrast, the availability-by-recall mechanism (78% correct predictions) clearly outperformed the other two mechanisms, whereas the fluency mechanism (69.8%) performed about seven percentage points better than the regressed-frequency mechanism (62.8%).

As pointed out, the mechanisms' discrimination rates (see Figure 4) differed extremely. To level the playing field, we next turned to a different kind of analysis. Specifically, we 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, we 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 1 (see the Predictions section, above).[11] Such an analysis would have the additional benefit of allowing us to compare results across studies. We used the same goodness-of-fit

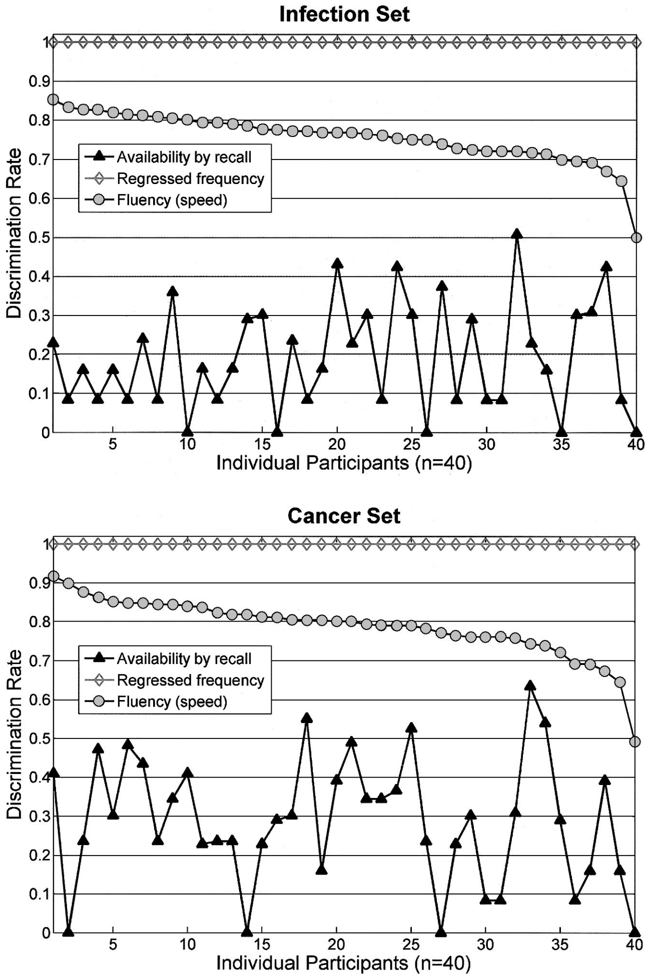

Figure 4. How often did the mechanisms make a prediction? Discrimination rates (for each of the 40 participants) for the availability-by-recall, regressed-frequency, and fluency mechanisms for the infection set (top) and the cancer set (bottom).

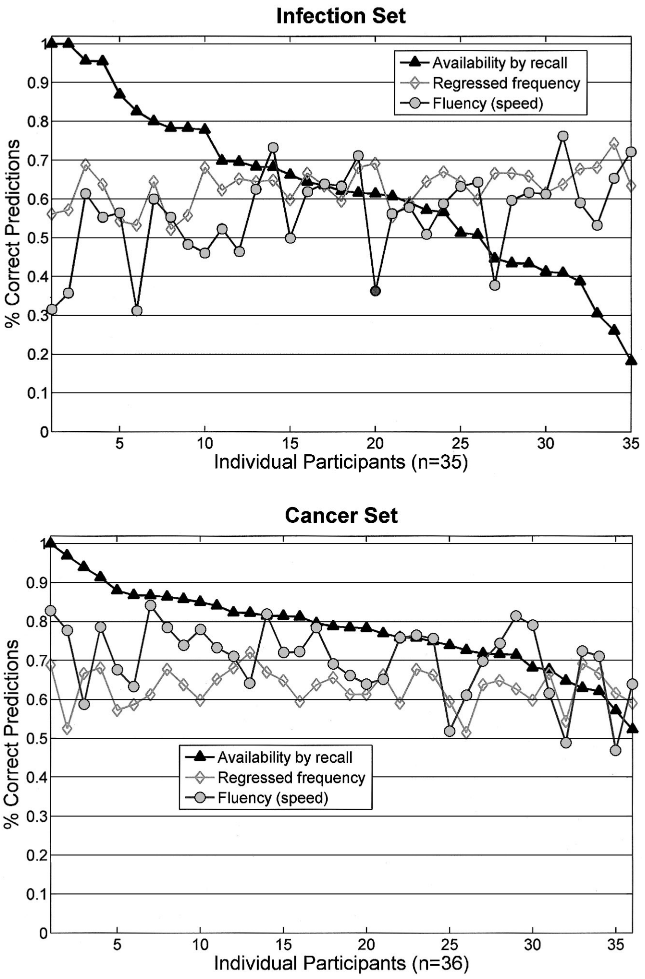

Figure 5. How often did the mechanisms make the correct prediction? Proportions of correctly predicted actual choices (within the set of comparisons in which a mechanism discriminated) for the infection set (top) and the cancer set (bottom).

[11] Both definitions of fluency were used. For the definition in terms of recognition speed, we 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).

criteria as in Study 1. As Figure 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). Thus, by and large, the analysis on the aggregate level mirrored the results obtained for individual responses.

Summary of Studies 1 and 2

We conducted two studies with a total of about 30,000 individual choices. In Study 1, we defined the notion of fluency 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 and Figure 3). In Study 2, we defined fluency 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—we 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 5, bottom panel) and was close to the fluency mechanism (media) when tested on the aggregated level (see Table 3 and Figure 3).

On the basis of Studies 1 and 2, we conclude that regardless of whether fluency is defined in terms of word frequency or recognition speed, its predictive power is limited. 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.[12] 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 3: 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 feel really safe, sometimes they would like to know how few people's lives have actually been claimed by plane crashes. Often, such a question comes to mind just after they have buckled themselves into an airplane seat. In this and many other situations, all they 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. A person cannot simply take the recalled number of deaths from, say, breast cancer (experienced in the person's social circle) as an estimate of the population mortality rate. Instead, he or she would need to, for instance, estimate how large his or her social circle is in relation to the total population and then adjust the 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, we describe how we tested which of the candidate mechanisms could account for mapping properties of frequency estimates.

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 1). 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 1; 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 1, botulism (in the infection set) was excluded from the final analysis (see footnote 9, above).

Results

Before we turn to the candidate mechanisms, let us describe the estimates and their accuracy in more detail. The median estimates for the three risk sets are reported in Table 1 (median estimates as they are not unduly influenced by outliers). Figure

[12] That is, they both had the same Spearman rank correlation in the assorted set in Study 1 (see Table 3).

6 shows the median estimates plotted against the actual frequencies in the assorted set. As did Lichtenstein et al. (1978), we observed what seems like overestimation of rare risks and underestimation of common risks.

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:

$$|\log_{10}(\text{estimated value/true value})|.$$

Table 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 our results 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. Averaged across all three sets, we found exactly the same mean absolute OME (see Table 4). This suggests that estimates of health risk frequencies are as accurate as estimates in other knowledge domains. Moreover, we found that accuracy 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 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 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 sets, 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, we 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.

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. We therefore examined the mechanisms' ability to predict to what extent the estimated values

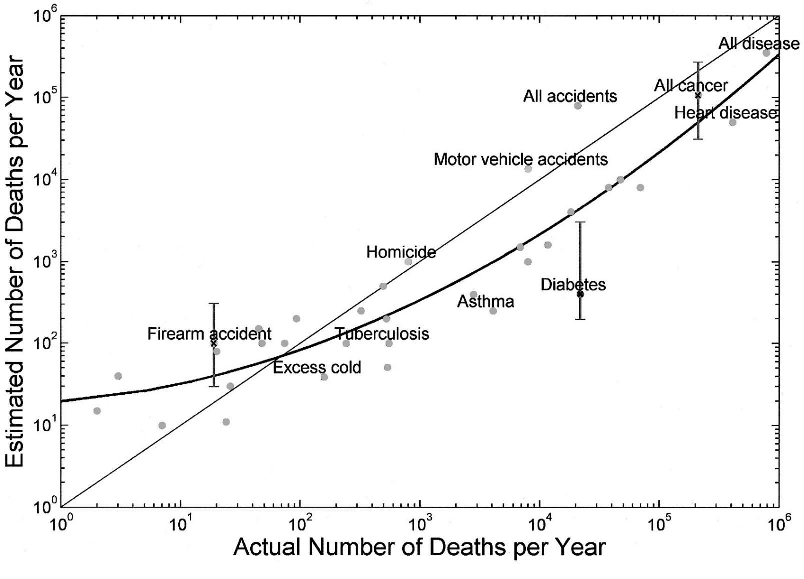

Figure 6. Estimates of risk frequencies: relationship between estimated and actual number of deaths per year for 41 causes of death in the assorted set. Each point represents the median estimate of 45 participants. The curved line is the best fitting quadratic regression line: Log median = 1.291 + 0.118 × log actual frequency + 0.098 × log actual frequency^2. Vertical bars depict the 25th and 75th percentiles of individual estimates for firearm accident, diabetes, and all cancer.

Table 4
Order of Magnitude Error (OME; Mean Absolute OME, Standard Error), Rank Correlation Between Median Estimated and Actual Frequencies (r_s), and Median of the Individual Rank Correlations (and Their Ranges) 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

followed the predicted values monotonically.[13] We used contrast analysis as the measure for the covariation of predictions and estimates (Rosenthal & Rosnow, 1985; Sedlmeier et al., 1998). Table 5 shows the results of the contrast analysis (MS_{contrast}, MS_{error}, df_{error}, and F value).

Table 5 also shows the effect size r associated with the four mechanisms (Rosenthal & Rosnow, 1985).[14] 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 1 and 2), 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 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.

So how did the mechanisms perform here? As was the case for judgments of which of two risks is more frequent (Studies 1 and 2), 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, we describe the main results and discuss their implications. Furthermore, we use the data we obtained to reconsider some of the conclusions drawn by Lichtenstein et al. (1978).

What We Have Learned

We 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 5 and Figure 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, we linked ease with the notion of fluency. We measured fluency 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, we 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. Our results, 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 1 and 2, 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 3, 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 9) across all three sets. This comparison resulted in a weighted (by *df*) mean effect size of

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

[14] 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 $L^2/n\Sigma\lambda^2$, 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.

Table 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 analysis of variance (instead of a between-groups analysis of variance; 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}. We did not test the risk-category mechanism in the cancer and infection sets because it would have predicted that each element within a set would receive the same estimate.

$r = .0001$ (see Table 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 1 and Study 2, the correlations are significant in all sets, they are far from perfect, that is, *r*s = .71, .87 (Study 2: .62), and .41 (Study 2: .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 1 and 2, 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 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.

We 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,

Table 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 3. 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 $r = [F/(F + df_{\text{error}})]^{1/2}$ (e.g., Rosenthal & Rosnow, 1991).

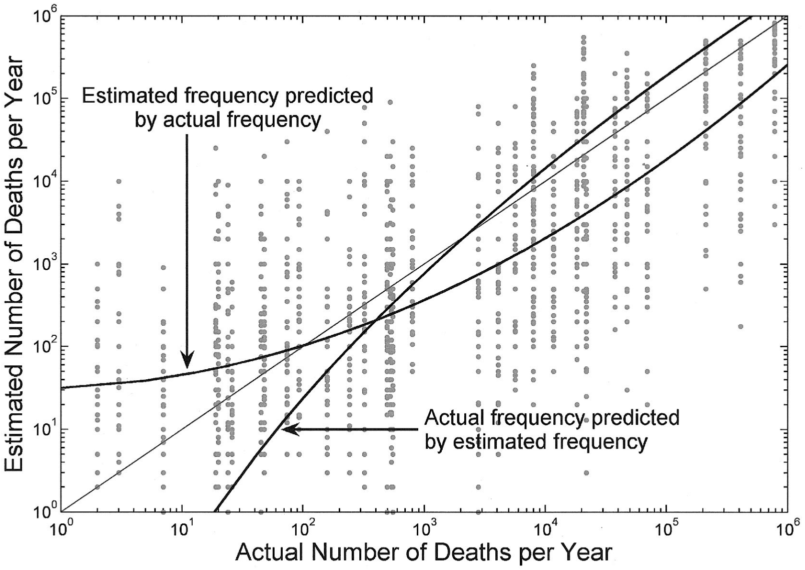

Figure 7. The primary bias and its reversal. The data and both best fitting quadratic regression lines are shown. Log estimated frequency = 1.5 + 0.057 × log actual frequency + 0.099 × log actual frequency^2. Log actual frequency = 1.27 + 0.47 × log estimated frequency + 0.045 × log estimated frequency^2. One regression line suggests the primary bias; the other regression line suggests its reversal.

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 our analysis.

Are Misjudgments the Result of a Systematic Bias?

Slovic et al. (1982) summarized the results of the Lichtenstein et al. (1978) study as follows: "Judgments were moderately accurate in a global sense: People usually knew which were the most and least frequently lethal events. Within this global picture, however, people made serious misjudgments, many of which seemed to reflect the influence of availability" (Slovic et al., 1982, p. 466).

Many later authors have taken Lichtenstein et al.'s (1978) results to show that people are plainly incapable of accurately judging risk frequencies. One representative voice is that of Sunstein (2002), who asked, "Do people know which risks lead to many deaths and which risks lead to few? They do not. In fact, they make huge blunders" (p. 1126; for similar conclusions, see Baron, 2000; Fischhoff, Bostrom, & Quadrel, 1993; Frost, Frank, & Maibach, 1997; Harding, Eiser, & Kristiansen, 1982; Lundborg & Lindgren, 2002; Toth & Daniels, 2002). Yet deviations from correct response can occur through factors other than systematic biases. In the following, we consider one such source and explore to what extent the inaccuracies we observed in Studies 1 and 3 can be accounted for by unsystematic variance.

One of the misjudgments that Lichtenstein et al. (1978) identified was the primary bias, according to which rare causes of death are overestimated and common causes of death are underestimated. Recently, Gigerenzer and Fiedler (2004) have suggested that the pattern of over- and underestimation as displayed in Figure 1 can be deduced from the existence of unsystematic conditional variability and that the assumption of a systematic bias would be unnecessary. Instead, the pattern can be explained as a consequence of regression toward the mean (for a similar argument in the context of overconfidence research, see Erev et al., 1994; for a debate on their argument, see Brenner, 2000, and Wallsten, Erev, & Budescu, 2000).

The crucial implication of this argument (a more detailed depiction of Gigerenzer and Fiedler's, 2004, point can be found in the Appendix) is that the specific pattern of over- and underestimation found should depend on how the data are grouped or conditioned. In Figure 7, two regression lines are plotted for the estimates of the assorted set in Study 3. When one predicts the estimated data from the actual frequencies—the first regression curve—it seems that rare risks are overestimated and common risks are underestimated, the original primary bias.[15] In contrast, when one predicts the actual data from the estimated frequencies—the second regression curve—the pattern is reversed. Now, rare risks are underestimated, and common risks are overestimated, the opposite of the primary bias. In other words, as a function of different conditioning and plotting of the data, one observes mutually exclusive phenomena.

A second consequence of unsystematic variance is that there is an expected proportion of errors (the less frequent risk judged to be the more frequent risk) in the choice task, again without having to

[15] Note that the regression line looks slightly different from that in Figure 4, as now individual estimates, rather than the median estimates, are being used.

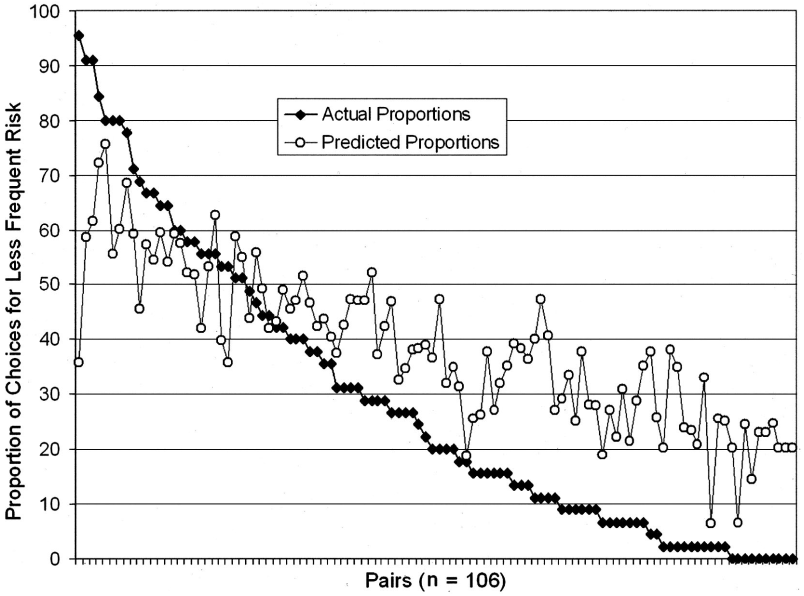

Figure 8. Incorrect choice and unsystematic variance: actual and predicted proportions of participants who selected the less frequent cause of death (assorted set) to be the more frequent one.

assume a systematic bias. To explore to what extent unsystematic variance could account for the errors we obtained for the choices in the assorted set in Study 1, we took advantage of Study 3's estimates and their interindividual variability.[16] Specifically, we used them to predict the probability that the rarer risk would be falsely considered to be more frequent (see the Appendix for how we determined the expected proportion of choices for the less frequent risk). Figure 8 plots for all 106 pairs in Study 1's assorted set the predicted and the actual proportions of people who falsely judged the rarer risk to be the more frequent risk. The Pearson correlation between predicted and actual proportions amounts to $r = .84$. In addition, for 77% (82 of the 106 pairs) of all comparisons, the predicted proportion of participants judging the less frequent cause of death to be more frequent was even larger than the actual proportion, whereas for 23% (18 of 106), the actual proportion was larger.

To conclude, a substantial proportion of the deviation between the environmental and the judged frequencies can be explained in terms of unsystematic variance. Admittedly, this variance did not drive all of the error in people's judgments. For instance, the two resulting regression lines in Figure 7 are not completely symmetrical to the identity line, as would be expected if the phenomenon were to be totally accounted for by regression toward the mean. Yet even this lack of complete symmetry need not indicate a cognitive bias. It may simply reflect a tendency not to use very high numbers in estimates of quantities. Alternatively, it may result from inferring actual rates through a shrinkage estimation procedure, which, in turn, has a Bayesian justification (see Stigler, 1990).

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, we are 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 we have found: The Pearson correlations (Spearman rank correlations) between the number of recalled cases and their actual frequencies in Study 1 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 2, 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? We suggest that one reason is the space in memory that the availability-by-recall mechanism can

[16] We are grateful to Thomas Wallsten, who suggested this analysis to us.

search. By requiring participants to recall personally experienced instances of death and illness, we defined the search space 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.[17] In contrast, sampling within one's social circle guards 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.

The Impact of Financial Incentives

Researchers have drawn far-reaching conclusions about people's lack of competence to judge the likelihood of risks. In light of these conclusions, a surprising divergence between the experimental practices of psychologists working in the field of behavioral decision making and those of experimental economists becomes relevant (see Hertwig & Ortmann, 2001, 2003). The latter treat the use of financial incentives as de rigueur (see Camerer & Hogarth, 1999). Arguably, the most important reason for economists' strict norm is their belief that if nothing is at stake in an experimental setting, participants may not bother to think carefully about the problem and therefore may respond in an offhand, unreliable fashion. In contrast, if appropriately paid, people's performance data are more likely to converge toward the performance criteria and are less variable (Smith & Walker, 1993). To the extent that this argument applies to judgments of risk frequencies, it provides another reason for why people's judgments of risk frequencies, at least in the eyes of some researchers, are quite poor—not because of lack of competence but because of lack of cognitive effort.

Only a few studies have examined the impact of financial incentives on the accuracy of risk frequency judgments (e.g., Harrison & Rutström, in press).[18] When evidence regarding the impact of financial incentives is lacking, Hertwig and Ortmann (2001) recommended that researchers use a simple do-it-both-ways rule, that is, they should examine incentive and nonincentive conditions, thus contributing to a growing knowledge of when and why incentives affect performance. Across all three studies, we examined people's risk judgments in incentive and no-incentive conditions and consistently obtained the same result: People's choices and estimates were not altered as a function of the presence or absence of financial incentives. As Tables 7 and 8 show, there were no marked differences as a function of incentives—neither in terms of measures of accuracy nor in terms of measures of variability. Using a ranking task (participants ranked various causes of death according to their actual frequencies), Harrison and Rutström (in press) also found no discernible difference between rankings with and without incentives. In our studies, in addition, financial incentives did not affect which candidate mechanisms explained people's choices best: As Table 7 shows, the magnitudes of the RMSDs between data and predictions were surprisingly similar in the incentive and no-incentive conditions.

Why did financial incentives not matter? Currently, we can only speculate about possible reasons. Here are three. First, retrieval processes involving recall and recognition may not require much cognitive effort (a view for which there is a growing body of evidence; e.g., Craik, Naveh-Benjamin, Ishaik, & Anderson, 2000). Second, the payoff decrement that participants accepted by not investing cognitive effort may have been too small to be considered meaningful. Although we cannot exclude this possibility, we took measures to ensure that people did not perceive the chance of getting financial incentives to be too slim to bother trying (in groups of four people, three received incentives). In addition, the income difference between the person who performed best and the one who performed worst was significant (i.e., amounted to 30% of the total payment). A final reason relates to the robust observation that simple choice and estimation strategies can do surprisingly well and even compete with strategies that demand more computation and information (Gigerenzer, Todd, & the ABC Research Group, 1999; Hertwig, Hoffrage, & Martignon, 1999). If so, then people who invest more cognitive

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

[18] However, there are studies that have examined the accuracy of frequency judgments as a function of incentives. Wright and Aboul-Ezz (1988), for instance, found that the squared error (between actual vs. estimated frequency) was lower in the incentive condition compared with a condition in which students received a flat fee only. The difference was of medium to large effect size ($\eta = .38$; see J. Cohen, 1988). Eta is identical to the Pearson product–moment correlation coefficient when $df = 1$.

Table 7
Accuracy (Percentage Correct) of Choices and Performance of Mechanisms With and Without Incentives

	Study 1		Study 2	
Accuracy/performance	Incentive (n = 50)	No incentive (n = 60)	Incentive (n = 40)	No incentive (n = 40)
Percentage correct				
M	69.6	70.3	62.5	62.2
Mdn	71.1	71.1	63.8	63.4
Range	55.6–80.0	48.9–82.2	48.2–74.3	52.2–69.6
SD	6.3	7.6	5.9	4.7
RMSD (predicted–actual choice proportions)				
Regressed frequency	22.1	20.4	22.3	23.7
Availability by recall	22.5	22.1	21.1	19.9
Fluency (media)	30.0	29.5	28.2	22.8
Risk category	32.2	30.8	27.4	27.8

Note. RMSD = root-mean-square deviation.

effort (due to financial incentives) may not achieve a higher level of accuracy than decision makers who rely on computationally much humbler strategies. From this view, more cognitive effort may not necessarily result in more accurate outcomes (Hertwig & Todd, 2004).

Conclusion

If indeed humankind is about to enter, as Glasser (2004) conjectured, 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. We see our investigations as another step toward developing more precise models of the cognitive underpinning of inferences about the environmental statistics of risks. We also hope to have advanced the debate over whether people's imperfect judgments reflect systematically biased information processing or are the natural consequence of their uncertainty about environmental statistics.

Table 8
Accuracy of Estimates in the Incentive and No-Incentive Conditions and Collapsed Across Both Conditions

Accuracy measure	Collapsed (N = 164)	Incentive (n = 80)	No incentive (n = 84)
OME (*Mdn*)	−0.005	−0.030	0.020
r_s (*Mdn*)	.50	.51	.49
Range	−.15–.92	−.06–.87	−.15–.92

Note. The data include a total of 3,725 observations (i.e., we collapsed estimates across the assorted set, the cancer set, and the infection set). OME: order of magnitude error.

References

Arbeitsgemeinschaft Bevölkerungsbezogener Krebsregister in Deutschland. (1999). *Krebs in Deutschland: Häufigkeiten und Trends* (2, aktualisierte Ausgabe). Saarbrücken, Germany: Author.

Baron, J. (2000). *Thinking and deciding* (3rd ed.). Cambridge, England: Cambridge University Press.

Barsalou, L. W. (1992). *Cognitive psychology: An overview for cognitive scientists*. Hillsdale, NJ: Erlbaum.

Begg, I., Maxwell, D., Mitterer, J. O., & Harris, G. (1986). Estimates of frequency: Attribute or attribution? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 496–508.

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

Benjamin, D. K., & Dougan, W. R. (1997). Individuals' estimates of the risks of death: Part I—A reassessment of the previous evidence. *Journal of Risk and Uncertainty, 15,* 115–133.

Benjamin, D. K., Dougan, W. R., & Buschena, D. (2001). Individuals' estimates of the risks of death: Part II—New evidence. *Journal of Risk and Uncertainty, 22,* 35–57.

Betsch, T., & Pohl, D. (2002). Tversky and Kahneman's availability approach to frequency judgement: A critical analysis. In P. Sedlmeier & T. Betsch (Eds.), *Etc.: Frequency processing and cognition* (pp. 109–119). Oxford, England: Oxford University Press.

Betsch, T., Siebler, F., Marz, P., Hormuth, S., & Dickenberger, D. (1999). The moderating role of category salience and category focus in judgments of set size and frequency of occurrence. *Personality and Social Psychology Bulletin, 25,* 463–481.

Brenner, L. (2000). Should observed overconfidence be dismissed as a statistical artifact? Critique of Erev, Wallsten, & Budescu (1994). *Psychological Review, 107,* 943–946.

Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1539–1553.

Brown, N. R. (2002a). Encoding, representing, and estimating event frequencies: A multiple strategy perspective. In P. Sedlmeier & T. Betsch (Eds.), *Etc.: Frequency processing and cognition* (pp. 37–53). Oxford, England: Oxford University Press.

Brown, N. R. (2002b). Real-world estimation: Estimation modes and seeding effects. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 41, pp. 321–360). New York: Academic Press.

Brown, N. R., Cui, X., & Gordon, R. D. (2002). Estimating national populations: Cross-cultural differences and availability effects. *Applied Cognitive Psychology, 16,* 811–827.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review, 100,* 511–534.

Burger, E. J., Jr. (1984). *Health risks: The challenge of informing the public*. Washington, DC: Media Institute.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19,* 7–42.

Carnegie Mellon University Graduate Research Methods Class. (1983). On judging the frequency of lethal events: A replication. *Risk Analysis, 3,* 11–16.

Christensen-Szalanski, J. J. J., Beck, D. E., Christensen-Szalanski, C. M., & Koepsell, T. D. (1983). Effects of expertise and experience on risk judgments. *Journal of Applied Psychology, 68,* 278–284.

Cohen, B. H. (1966). Some-or-none characteristics of coding behavior. *Journal of Verbal Learning and Verbal Behaviour, 5,* 182–187.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Combs, B., & Slovic, P. (1979). Newspaper coverage of causes of death. *Journalism Quarterly, 56,* 837–843.

Craik, F. I. M., Naveh-Benjamin, M., Ishaik, G., & Anderson, N. D. (2000). Divided attention during encoding and retrieval: Differential control effects? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1744–1749.

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General, 130,* 579–599.

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

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519–527.

Fiedler, K. (1983). On the testability of the availability heuristic. In R. W. Scholz (Ed.), *Decision making under uncertainty: Cognitive decision research, social interaction, development and epistemology* (pp. 109–119). Amsterdam: North-Holland.

Fiedler, K., & Armbruster, T. (1994). Two halves may be more than one whole: Category-split effects on frequency illusions. *Journal of Personality and Social Psychology, 66,* 633–645.

Fischhoff, B., Bostrom, A., & Quadrel, M. J. (1993). Risk perception and communication. *Annual Review of Public Health, 14,* 183–203.

Fiske, S. T., & Pavelchak, M. A. (1986). Category-based versus piecemeal-based affective responses: Developments in schema-triggered affect. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (pp. 167–203). New York: Guilford Press.

Folkes, V. S. (1988). The availability heuristic and perceived risk. *Journal of Consumer Research, 15,* 13–23.

Fraisse, P. (1984). Perception and estimation of time. *Review of Psychology, 35,* 1–36.

Frost, K., Frank, E., & Maibach, E. (1997). Relative risk in the news media: A quantification of misrepresentation. *American Journal of Public Health, 87,* 842–845.

Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology, 8,* 172–179.

Gigerenzer, G., & Fiedler, K. (2004). *Minds in environments: The potential of an ecological approach to cognition*. Manuscript submitted for publication.

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

Glasser, R. J. (2004, July). We are not immune: Influenza, SARS, and the collapse of public health. *Harper's Magazine, 309*(1850), 35–42.

Greene, R. L. (1984). Incidental learning of event frequency. *Memory & Cognition, 12,* 90–95.

Harding, C. M., Eiser, J. R., & Kristiansen, C. M. (1982). The representation of mortality statistics and the perceived importance of causes of death. *Journal of Applied Social Psychology, 12,* 169–181.

Harrison, G. W., & Rutström, E. E. (in press). Eliciting subjective beliefs about mortality risk orderings. *Environmental & Resource Economics.*

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108,* 356–388.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39,* 1372–1388.

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

Hertwig, R., Hoffrage, U., & Martignon, L. (1999). Quick estimation: Letting the environment do the work. In G. Gigerenzer, P. M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (pp. 209–234). New York: Oxford University Press.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists. *Behavioral and Brain Sciences, 24,* 383–451.

Hertwig, R., & Ortmann, A. (2003). Economists' and psychologists' experimental practices: How they differ, why they differ and how they could converge. In I. Brocas & J. D. Carillo (Eds.), *The psychology of economic decisions* (pp. 253–272). New York: Oxford University Press.

Hertwig, R., & Todd, P. M. (2004). More is not always better: The benefits of cognitive limits. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment and decision making* (pp. 213–232). Chichester, England: Wiley.

Hintzman, D. L. (1969). Apparent frequency as a function of frequency and the spacing of repetitions. *Journal of Experimental Psychology, 80,* 139–145.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace model. *Psychological Review, 95,* 528–551.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129,* 220–241.

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. *Psychology of Learning and Motivation, 18,* 1–47.

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

Jonides, J., & Jones, C. M. (1992). Direct coding for frequency of occurrence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 368–378.

Kristiansen, C. M. (1983). Newspaper coverage of diseases and actual mortality statistics. *European Journal of Social Psychology, 13,* 193–194.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 551–578.

Lundborg, P., & Lindgren, B. (2002). Risk perceptions and alcohol consumption among young people. *Journal of Risk and Uncertainty, 25,* 165–183.

MacLeod, C., & Campbell, L. (1992). Memory accessibility and probability judgments: An experimental evaluation of the availability heuristic. *Journal of Personality and Social Psychology, 63,* 890–902.

Meyer, P. (1990). News media responsiveness to public health. In C. Atkin & L. Wallack (Eds.), *Mass communication and public health* (pp. 52–59). Newbury Park, CA: Sage.

Nickerson, R. (1981). Motivated retrieval from archival memory. In H. E.

Howe, Jr., & J. H. Flowers (Eds.), *Nebraska Symposium on Motivation: Vol. 28. Cognitive processes* (pp. 73–119). Lincoln: University of Nebraska Press.

Oppenheimer, D. (2004). Spontaneous discounting of availability in frequency judgment tasks. *Psychological Science, 15,* 100–105.

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

Pennington, H. (2004, July 8). Why can't doctors be more scientific? *London Review of Books, 26*(13), 28–29.

Robert Koch Institute. (2001). *Epidemiologisches Bulletin, 20/2001*. Berlin, Germany: Author.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge, England: Cambridge University Press.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Rosnow, R. L., & Rosenthal, R. (1996). Contrasts and interactional redux: Five easy pieces. *Psychological Science, 7,* 253–257.

Ross, S. M. (2000). *Introduction to probability models* (7th ed.). San Diego, CA: Academic Press.

Schooler, L. J., & Hertwig, R. (in press). How forgetting aids heuristic inference. *Psychological Review.*

Schwarz, N., & Vaughn, L. A. (2002). The availability heuristic revisited: Ease of recall and content as distinct sources of information. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 103–119). Cambridge, England: Cambridge University Press.

Schwarz, N., & Wänke, M. (2002). Experiential and contextual heuristics in frequency judgement: Ease of recall and response scales: In P. Sedlmeier & T. Betsch (Eds.), *Etc.: Frequency processing and cognition* (pp. 89–108). Oxford, England: Oxford University Press.

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

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 48*(A), 257–279.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1982). Facts versus fears: Understanding perceived risk. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 463–489). Cambridge, England: Cambridge University Press.

Smith, V. L., & Walker, J. M. (1993). Monetary rewards and decision costs in experimental economics. *Economic Inquiry, 31,* 245–261.

Stapel, D. A., Reicher, S. D., & Spears, R. (1994). Social identity, availability and the perception of risk. *Social Cognition, 12,* 1–17.

Statistisches Bundesamt. (2002). *Gesundheitswesen: Sterbefälle nach Todesursachen in Deutschland, Einzelnachweis (ICD-10), 2000.* Wiesbaden, Germany: Author.

Stigler, S. M. (1990). A Galtonian perspective on shrinkage estimators. *Statistical Science, 5,* 147–155.

Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.

Sunstein, C. R. (2002). The laws of fear: Review of *The Perception of Risk,* by Paul Slovic. *Harvard Law Review, 115,* 1119–1168.

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

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

Tversky, A., & Kahneman, D. (1974, September 27). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Wallsten, D., Erev, I., & Budescu, D. (2000). The importance of theory: Response to Brenner (2000). *Psychological Review, 107,* 947–949.

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

Williams, K., & Durso, F. T. (1986). Judging category frequency: Automaticity or availability? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 387–396.

World Health Organization. (1992). *ICD-10: International Statistical Classification of Diseases and Related Health Problems*. Geneva, Switzerland: Author.

Wright, W. F., & Aboul-Ezz, M. E. (1988). Effects of extrinsic incentives on the quality of frequency assessments. *Organizational Behavior and Human Decision Processes, 41,* 143–152.

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Etc.: Frequency processing and cognition* (pp. 21–36). Oxford, England: Oxford University Press.

Zanuck, R. D. (Producer), Brown, D. (Producer), & Spielberg, S. (Director). (1975). *Jaws* [Motion picture]. United States: Universal Pictures.

(*Appendix follows*)

Appendix

The Primary Bias Explained in Terms of Regression Toward the Mean (Gigerenzer & Fiedler, 2004)

Assume two variables X (e.g., actual frequencies) and Y (estimated mortality rates) that are bivariately normally distributed and have zero means, common variance σ^2, and correlation ρ (see Furby, 1973; Stigler, 1999). The expectation of Y given $X = x$ is

$$E(Y|X = x) = \rho \mathrm{x}, \tag{1}$$

and the expectation of X given $Y = y$ is

$$E(X|Y = y) = \rho \mathrm{y}. \tag{2}$$

Only if ρ is perfect (i.e., –1.0 or 1.0) will there be no regression toward the mean. Otherwise, the expected value (e.g., Y) will be closer to the mean than the predictor (e.g., x). Furthermore, because the conditional variance (e.g., the variance of Y given $X = x$) is related to ρ by

$$Var(Y|X = x) = Var(X|Y = y) = (1 - \rho^2)\sigma^2, \tag{3}$$

it follows that if the conditional variance is larger than zero, ρ will be imperfect, and regression toward the mean will occur. It is important to note that a less than perfect correlation between X and Y could be, for instance, due to unsystematic cognitive factors such as information loss during memory storage (Fiedler & Armbruster, 1994) or due to environmental factors (i.e., ups and downs in death statistics across time).

Calculation of the Expected Errors in the Choice Task

Assuming the estimates for the more frequent Risk a and the less frequent Risk b as random variables that are independent[A1] and normally distributed, we can also represent the difference between the distributions of a and b as a normally distributed random variable (Ross, 2000, p. 68), $N(e_a - e_b, \sigma^2{}_a + \sigma^2{}_b)$, with $M = (e_a - e_b)$ and

$$SD = \sqrt{\sigma_a^2 + \sigma_b^2}.$$

The probability that the rarer Event b is falsely judged to be more frequent than Event a equals the probability that the standard normal distribution $N(0, 1)$ obtains a value smaller than

$$\phi = \frac{-\,e_a + e_b}{\sqrt{\sigma_a^2 + \sigma_b^2}}.$$

[A1] Admittedly, this is a simplifying assumption, as each participant provided estimates for all events.

Received June 27, 2003
Revision received December 17, 2004
Accepted January 4, 2005 ■