Statistical Thinking: No One Left Behind

Björn Meder and Gerd Gigerenzer

Abstract Is the mind an "intuitive statistician"? Or are humans biased and error-prone when it comes to probabilistic thinking? While researchers in the 1950s and 1960s suggested that people reason approximately in accordance with the laws of probability theory, research conducted in the heuristics-and-biases program during the 1970s and 1980s concluded the opposite. To overcome this striking contradiction, psychologists more recently began to identify and characterize the circumstances under which people—both children and adults—are capable of sound probabilistic thinking. One important insight from this line of research is the power of representation formats. For instance, information presented by means of natural frequencies, numerical or pictorial, fosters the understanding of statistical information and improves probabilistic reasoning, whereas conditional probabilities tend to impede understanding. We review this research and show how its findings have been used to design effective tools and teaching methods for helping people—be it children or adults, laypeople or experts—to reason better with statistical information. For example, using natural frequencies to convey statistical information helps people to perform better in Bayesian reasoning tasks, such as understanding the implications of diagnostic test results, or assessing the potential benefits and harms of medical treatments. Teaching statistical thinking should be an integral part of comprehensive education, to provide children and adults with the risk literacy needed to make better decisions in a changing and uncertain world.

B. Meder (✉) · G. Gigerenzer
Center for Adaptive Behavior and Cognition (ABC), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany
e-mail: meder@mpib-berlin.mpg.de

G. Gigerenzer
Harding Center for Risk Literacy, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

1 Introduction

Is the mind an "intuitive statistician"? Or are humans biased and error-prone when it comes to probabilistic thinking? Several studies have addressed these questions, giving very different answers. Some research suggests that, by and large, people's intuitive inferences approximately correspond to the laws of probability calculus, mirroring the Enlightenment view that the laws of probability are also the laws of the mind (Daston 1988). Other studies indicate that people have severe difficulties with probability judgments, which has led researchers to conclude that the human mind is not build to reason with probabilities (Tversky and Kahneman 1974).

Much of the work used to corroborate either of the two positions has focused on simple statistical inference problems. Consider a woman applying a home pregnancy test kit. If she tests positive, the probability that she is pregnant increases. Conversely, if the test is negative, the probability that she is pregnant decreases. However, because such tests are not perfectly reliable, these inferences are not certain, but can only be probabilistic. For instance, the sensitivity of home test kits (probability of obtaining a positive test given pregnancy) can be as low as 75 % when applied by inexperienced people (Bastian et al. 1998). This means that out of 100 women who are pregnant, only 75 would test positive, while 25 would test negative. At the same time, such tests are not perfectly specific, meaning that a woman who is not actually pregnant may nevertheless get a positive test result (Bastian et al. 1998).

Situations like this are called *Bayesian problems*, since probability theory, and Bayes' rule in particular, serve as a reference for examining how people revise their beliefs about some state of the world (e.g., being pregnant) in the light of new evidence (e.g., a test result). What are the determinants and limitations of sound reasoning in such tasks? We review research on probabilistic reasoning and show how its findings have been used to design effective tools and teaching methods for helping people—be it children or adults, laypeople or experts—to reason more appropriately with statistical information. Our discussion centers on one of psychologists' *drosophila* for investigating probabilistic thinking, an elementary form of Bayesian reasoning requiring an inference from a single, binary observation to a single, binary hypothesis. Many studies offer a pessimistic view on people's capacity to handle such problems, indicating that both John Q. Public and experts have severe difficulties with them (Kahneman and Tversky 1973; Tversky and Kahneman 1974). However, more recent studies have provided novel insights into the circumstances under which people take the normatively relevant variables into account and are able to solve such problems (Gigerenzer and Hoffrage 1995). Instead of emphasizing human errors, the focus is shifted to human engineering: What can (and need) be done to help people with probabilistic inferences?

One way to foster reasoning with statistical information is to convey information in a transparent and intuitive manner. For instance, a number of studies show that certain frequency formats strongly improve the reasoning of both laypeople in the laboratory (Cosmides and Tooby 1996; Gigerenzer and Hoffrage 1995) and experts outside it (Gigerenzer et al. 2007; Hoffrage and Gigerenzer 1998; Labarge et al.

2003). Drawing on these findings, effective methods have been developed to improve people's ability to reason better with statistical information (Sedlmeier 1999; Sedlmeier and Gigerenzer 2001). These studies show that sound probabilistic thinking is not a mysterious gift out of reach for ordinary people, but that one can learn to make better inferences by using the power of representation: "Solving a problem simply means representing it so as to make the solution transparent" (Simon 1969, p. 153).

2 Bayesian Reasoning as a Test Case of Probabilistic Thinking

In situations like the pregnancy test, a piece of evidence is used to revise one's opinion about a hypothesis. For instance, how does a positive test change the probability of being pregnant? From a statistical point of view, answering this question requires taking into account the prior probability of the hypothesis H (i.e., the probability of being pregnant *before* the test is applied) and the likelihood of obtaining datum D under each of the hypotheses (i.e., the likelihood of a positive test when pregnant and the likelihood of a positive test when not pregnant). From this, the posterior probability of being pregnant given a positive test, $P(H|D)$, can be computed by Bayes' rule:

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\neg H) \times P(\neg H)} = \frac{P(D|H) \times P(H)}{P(D)} \quad (1)$$

where $P(H)$ denotes the hypothesis' prior probability, $P(D|H)$ denotes the probability of observing datum D given the hypothesis is true, and $P(D|\neg H)$ denotes the probability of D given the hypothesis is false.

2.1 Is the Mind an "Intuitive Statistician"?

Bayes' rule is an uncontroversial consequence of the axioms of probability theory. Its status as a descriptive model of human thinking, however, is not at all self-evident. In the 1950s and 1960s, several studies were conducted to examine to what extent people's intuitive belief revision corresponds to Bayes' rule (e.g., Edwards 1968; Peterson and Miller 1965; Peterson et al. 1965; Phillips and Edwards 1966; Rouanet 1961). For instance, Phillips and Edwards (1966) presented subjects with a sequence of draws from a bag containing blue and red chips. The chip was drawn from either a predominantly red bag (e.g., 70 % red chips and 30 % blue chips) or a predominantly blue bag (e.g., 30 % red chips and 70 % blue chips). Given a subject's prior belief about whether it came from a predominantly red or predominantly blue bag, the question of interest was whether subjects would update their beliefs in accordance with Bayes' rule.

This and several other studies showed that participants took the observed evidence into account to some extent, but not as extensively as prescribed by Bayes'

rule. "It turns out that opinion change is very orderly and usually proportional to numbers calculated from Bayes's theorem—but it is insufficient in amount." (Edwards 1968, p. 18). Edwards and colleagues called this observation *conservatism*, meaning that people shifted their probability estimates in the right direction, but did not utilize the data as much as an idealized Bayesian observer would. While there was little disagreement on the robustness of the phenomenon, different explanations were put forward (see Edwards 1968, for a detailed discussion). One idea was that conservatism results from a *misaggregation* of information, such as a distorted integration of priors and likelihoods. Other researchers suggested that the inference process itself principally follows Bayes' rule, but a *misperception* of the data-generating processes and the diagnostic value of data would result in estimates that are too conservative. Another idea was that the predominantly used book, bag, and poker chip tasks were too artificial to draw more general conclusions and that people outside the laboratory would have less difficulty with such inferences.

Despite some systematic discrepancies between Bayes' rule and people's inferences, the human mind was considered to be an (albeit imperfect) "intuitive statistician": "Experiments that have compared human inferences with those of statistical man show that the normative model provides a good first approximation for a psychological theory of inference." (Peterson and Beach 1967, p. 43).

2.2 Is the Human Mind Biased and Error-Prone when It Comes to Probabilistic Thinking?

Only a few years later, other researchers arrived at a very different conclusion: "In making predictions and judgments under uncertainty, people do not appear to follow the calculus of chance or the statistical theory of prediction." (Kahneman and Tversky 1973, p. 237).

What had happened? Other studies had been conducted, also using probability theory as a normative (and potentially descriptive) framework. However, people's behavior in these studies seemed to be error-prone and systematically biased (Bar-Hillel 1980; Tversky and Kahneman 1974; Kahneman and Tversky 1972, 1973; Kahneman et al. 1982).

One (in)famous example of such an inference task is the so-called "mammography problem" (adapted from Eddy 1982; see Gigerenzer and Hoffrage 1995):

> The probability of breast cancer is 1 % for a woman at the age of 40 who participates in routine screening. If a woman has breast cancer, the probability is 80 % that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6 % that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? ___%

Here, the hypothesis in question is whether the woman has breast cancer given the base rate of cancer (1 %) and the likelihood of obtaining a positive test result for women with cancer (80 %) or without cancer (9.6 %).

From the perspective of statistical inference, the problem is simple. The relevant pieces of information are the prior probability of disease, $P(\text{cancer})$, the probability of obtaining a positive test result for a woman having cancer, $P(T+|\text{cancer})$, and the probability of obtaining a positive test result for a woman having *no* cancer, $P(T+|\text{no cancer})$. From this information, the probability of breast cancer given a positive test result, $P(\text{cancer}|T+)$, can be easily computed by Bayes' rule:

$$P(\text{cancer}|T+) = \frac{P(T+|\text{cancer}) \times P(\text{cancer})}{P(T+)}$$

$$= \frac{P(T+|\text{cancer}) \times P(\text{cancer})}{P(T+|\text{cancer}) \times P(\text{cancer}) + P(T+|\text{no cancer}) \times P(\text{no cancer})}$$

$$= \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.096 \times 0.99} = 0.078 \approx 8\,\%.$$

Thus, the probability that a woman with a positive mammogram has cancer is about 8 %. In stark contrast, empirical research shows that both health care providers and laypeople tend to give much higher estimates, often around 70 %–80 % (Casscells et al. 1978; Eddy 1982; Gigerenzer and Hoffrage 1995; Hammerton 1973; for reviews, see Bar-Hillel 1980; Koehler 1996a). These overestimates were interpreted as evidence for *base rate neglect*, suggesting that people did not seem to take base rate information (e.g., prior probability of cancer) into account to the extent prescribed by Bayes' rule.

These and similar findings were bad news for the notion of "statistical man" (Peterson and Beach 1967). After all, from a formal point of view, this problem is essentially as easy as it can get: A binary hypothesis space (cancer vs. no cancer), a single datum (positive test result), and all the information necessary to make the requested inference is available. If people failed to solve such an apparently simple problem, it seemed clear that humans lack the capacity to reason in accordance with Bayes' rule: "In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all." (Kahneman and Tversky 1972, p. 450).

3 The Power of Presentation Formats

While researchers in the 1950s and 1960s believed that people reason approximately in accordance with the laws of probability theory, albeit conservatively, studies conducted in the heuristics-and-biases program during the 1970s and 1980s concluded the opposite. To overcome this conceptual impasse, psychologists more recently began to identify and characterize the circumstances under which people—both children and adults—are capable of sound probabilistic thinking.

Gigerenzer and Hoffrage (1995) argued that one crucial component is the link between cognitive processes and information formats. For instance, mathematical operations like multiplication and division are hard with numbers represented as Roman numerals, but comparatively easy with Arabic numerals. The theoretical general point is that the representation does part of the job and can facilitate reasoning.

Gigerenzer and Hoffrage (1995; see also Cosmides and Tooby 1996) compared the above version of the mammography problem, in which numerical information is presented in terms of conditional probabilities, with a version in which the same information is expressed in terms of *natural frequencies*:

> 10 out of every 1,000 women at the age of 40 who participate in routine screening have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. Here is a new representative sample of women at the age of 40 who got a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer? ___ out of ___ .

When the problem was presented this way, a sharp increase in correct ("Bayesian") answers was obtained. For instance, in the mammography problem, 16 % of participants gave the correct solution when information was presented as probabilities, as opposed to 46 % in the natural frequency version. Similar results were obtained across 15 different problems (Gigerenzer and Hoffrage 1995).

Why is this? One general idea is that the human mind is better adapted to reason with frequency information, as this is the "raw data" we experience in our daily life and have adapted to throughout evolutionary history (Cosmides and Tooby 1996). For instance, over the course of professional life, a physician may have diagnosed hundreds of patients. The doctor may have experienced that out of 1,000 patients, few will actually have a certain disease (e.g., 10 out of 1,000) and that most of those (e.g., 8 out of 10) had a certain symptom. But the physician may also have experienced that of the many people who do not have the disease some also show the symptom (e.g., 95 out of the 990 without the disease). Now, if confronted with a new patient showing the symptom, the physician will know from past experience that only 8 out of 103 patients with the symptom actually have the disease. Thus, while aware of the fact that most people with the disease show the symptom, the physician is also well aware of the relatively high number of false positives (i.e., people without the disease but having the symptom) resulting from the low base rate of the disease (Christensen-Szalanski and Bushyhead 1981). If the relation between diagnostic cues and base rates is experienced this way, people's Bayesian reasoning has been shown to improve (Christensen-Szalanski and Beach 1982; see also Koehler 1996a; Meder and Nelson 2012; Meder et al. 2009; Nelson et al. 2010).

This leads directly to a second, more specific explanation: Natural frequencies facilitate probabilistic inference because they simplify the necessary calculations. A different way of computing a posterior probability according to Bayes' rule (Eq. (1)) is

$$P(D|H) = \frac{N(H \cap D)}{N(D)} \tag{2}$$

where $N(H \cap D)$ denotes the number of instances in which the hypothesis is true and the datum was observed (e.g., patients with disease and symptom) and $N(D)$ denotes the total number of cases in which D was observed (e.g., all patients with the symptom). (The equivalence of Eqs. (1) and (2) follows from the axioms of

probability theory, according to which $P(H \cap D) = P(D|H)P(H)$.) For the mammography problem, Eq. (2) yields

$$P(\text{cancer}|\text{positive test}) = \frac{N(\text{cancer} \cap \text{positive test})}{N(\text{positive test})} = \frac{8}{8+95} = 0.078 \approx 8\ \%$$

where $N(\text{cancer} \cap \text{positive test})$ denotes the number of women with cancer and a positive mammogram, and $N(\text{positive test})$ denotes the total number of cases with a positive test result.

Although this notation is mathematically equivalent to Eq. (1), it is not psychologically equivalent. Representing the probabilistic structure of the environment in terms of natural frequencies simplifies Bayesian reasoning because the required mathematical operation can be performed on natural numbers rather than on normalized fractions (i.e., probabilities). And because the base rate information is already contained in the natural frequencies that enter the computation, it must not be explicitly considered in the calculations. Thus, the representation does part of the job.

3.1 What Are Natural Frequencies and why Do They Simplify Bayesian Reasoning?

It is important to understand what natural frequencies are and what they are not. For instance, the findings of Gigerenzer and Hoffrage (1995) have sometimes been interpreted as suggesting that *any* type of frequency information facilitates probabilistic reasoning (e.g., Evans et al. 2000; Girotto and Gonzalez 2001; Johnson-Laird et al. 1999; Lewis and Keren 1999). However, the claim is that natural frequencies improve Bayesian reasoning (Gigerenzer and Hoffrage 1999; Hoffrage et al. 2002). By contrast, relative (normalized) frequencies should not—and did not—facilitate reasoning (Gigerenzer and Hoffrage 1995).

To understand why natural frequencies—but not frequencies in general—facilitate Bayesian reasoning, one needs to distinguish between *natural sampling* and *systematic sampling* (Kleiter 1994). The important difference is that natural samples preserve base rate information, whereas systematic sampling—the usual approach in scientific research—fixes base rates a priori. For example, if we take a random sample of 1,000 women from the general population, we will find that 10 of these 1,000 women have breast cancer (Fig. 1, left). Furthermore, we would observe that out of the 10 women with cancer, 8 will get a positive mammogram, and that 95 out of the 990 women without cancer will get a positive mammogram. The probability of women having cancer given a positive test result can now be easily calculated by comparing the (natural) frequency of women with cancer *and* a positive test result to the overall number of women with a positive mammogram (Fig. 1, left), namely, $8/(95+8)$ (Eq. (2)). As mentioned earlier, because these numbers already contain base rate information, it is unnecessary to explicitly consider this information when evaluating the implications of a positive test result.

Natural Frequency Tree   Normalized Frequency Tree
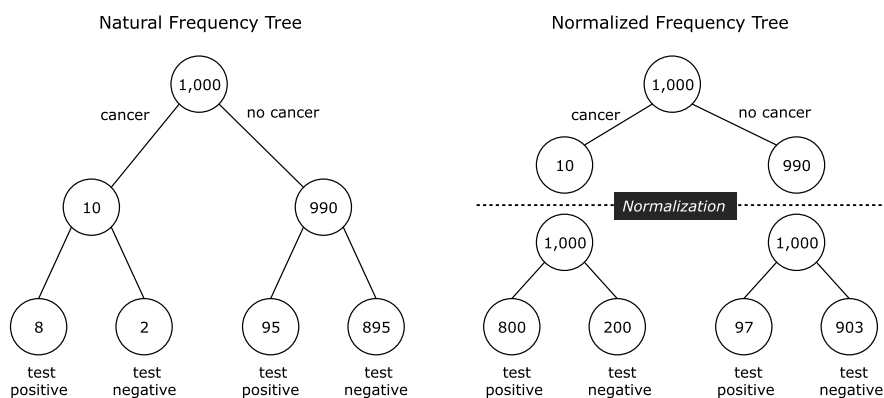
Fig. 1 A natural frequency tree (*left*) and a normalized frequency tree (*right*). The four numbers at the *bottom of the left tree* are natural frequencies; the four numbers at the *bottom of the right tree* are not. The natural frequency tree results from natural sampling, which preserves base rate information (number of women with cancer vs. without cancer in the population). In the normalized tree, the base rate of cancer in the population (10 out of 1,000) is normalized to an equal number of people with and without cancer (e.g., 1,000 women with and 1,000 women without cancer). In the natural frequency tree, the posterior probability of P(cancer | test positive) can be read from the tree by comparing the number of people with cancer and a positive test with the total number of positive test results ($8/(8 + 95)$). This does not work when base rates have been fixed a priori through systematic sampling (*right tree*)

However, this calculation is only valid with natural frequencies. It does not work with normalized (relative) frequencies resulting from systematic sampling, in which the base rate of an event in the population is normalized. For instance, to determine the specificity and sensitivity of a medical test, one might set up an experiment with an equal number of women with and without cancer (e.g., 1,000 women in each group) and apply the test to each woman (Fig. 1, right). The results make it possible to assess the number of true positives (women with cancer who get a positive test result) and false positives (women without cancer who get a positive test result) as well as the number of true negatives (women without cancer who get a negative test result) and false negatives (women with cancer who get a negative test result). This method is appropriate when evaluating test characteristics (e.g., sensitivity and specificity of a medical test) because it ensures that samples of women with and without cancer are large enough for statistically reliable conclusions to be drawn. However, when used to make diagnostic inferences, base rates must be explicitly reintroduced into the calculation (via Eq. (1)).

4 Can Children Solve Bayesian Problems?

The classic view endorsed by Piaget and Inhelder (1975) holds that young children lack the specific capacities necessary for sound probability judgments. For instance, the application of a combinatorial system and the capacity to calculate proportions

are prerequisites for reasoning with probabilities, abilities that are assumed not to develop before the age of 11 or 12.

On the other hand, studies indicate that even young children have basic intuitions regarding probability concepts (Fischbein et al. 1970; Girotto and Gonzalez 2008; Yost et al. 1962). Brainerd (1981) demonstrated some basic skills in children for understanding the outcomes of sampling processes given a certain distribution of events in a reference class. In one experiment, 10 tokens of two different colors were placed in an opaque container (e.g., 7 red and 3 black tokens). When children were asked to predict the outcome (e.g., black or red token) of a series of random draws, their predictions for the first draw usually preserved the ordering of the sampling probabilities. Although children's predictions were not consistent with the current sampling proportions in subsequent trials, when explicitly probed about the relative frequencies before each prediction, the performance of both younger children (preschoolers age 4 to 5) and older children (2nd and 3rd graders) improved (Brainerd 1981, Studies 6 and 12). The observed developmental trajectories also point to the link between probability judgments and basic cognitive processes, such as the storage, retrieval, and processing of frequency information in and from memory. A similar development across age groups was observed by Girotto and Gonzalez (2008), who showed that starting around the age of 5, children are sensitive to new evidence when betting on random events (i.e., draws from a bag containing chips) and can guess what outcome is more likely.

Using natural frequencies, can children also make more quantitative inferences in Bayesian reasoning tasks? Zhu and Gigerenzer (2006; see also Multmeier 2012) investigated children's (4th, 5th, and 6th graders) capacity to reason in accordance with Bayes' rule by using similar tasks to the mammography problem. One of these, for instance, was the "red nose" problem. This is the problem stated in terms of conditional probabilities (expressed as percentages):

> Pingping goes to a small village to ask for directions. In this village, the probability that the person he meets will lie is 10 %. If a person lies, the probability that he/she has a red nose is 80 %. If a person doesn't lie, the probability that he/she also has a red nose is 10 %. Imagine that Pingping meets someone in the village with a red nose. What is the probability that the person will lie?

The solution to this question can be inferred using Bayes' rule (Eq. (1)), which gives a posterior probability of 47 %.

Are children able to solve such a problem? The answer is no: None of the 4th, 5th, or 6th graders could solve such a problem when information was presented in terms of conditional probabilities. Of the adult control sample, about half of the participants were able to solve the problem (Fig. 2).

The main question was whether children would do better if the same information was presented in terms of natural frequencies. This is the same red nose problem in a natural frequency format:

> Pingping goes to a small village to ask for directions. In this village, 10 out of every 100 people will lie. Of the 10 people who lie, 8 have a red nose. Of the remaining 90 people who don't lie, 9 also have a red nose. Imagine that Pingping meets a group of people in the village with red noses. How many of these people will lie? ____ out of ____.

Fig. 2 Percentage of correct solutions in Zhu and Gigerenzer (2006; results aggregated across Studies 1 and 2). No child was able to solve a Bayesian reasoning task when information was presented probabilities. Performance was much better when reasoning with natural frequencies
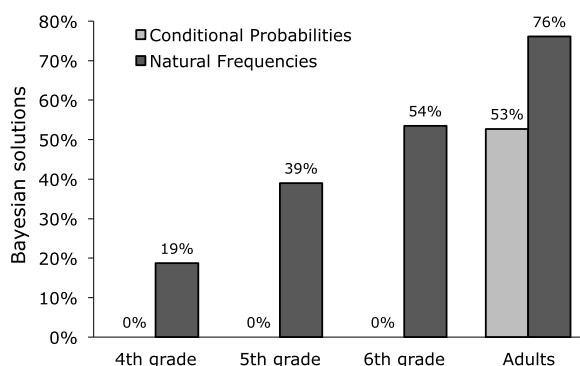
The findings show that natural frequencies can help children to understand and solve problems that are otherwise beyond their skills (Fig. 2). The results also show a strong trend with age, similar to results of other studies asking children for categorical probability judgments (Brainerd 1981; Girotto and Gonzalez 2008).

5 Bayesian Reasoning Outside the Laboratory

Apart from conducting laboratory studies on Bayesian reasoning (in which the participants are often students), researchers have also examined experts' capacity to reason in accordance with Bayes' rule. This is important, given that making inferences from statistical data is an important part of decision making in many areas, such as medicine and the law.

However, research shows that statistical illiteracy is a widespread phenomenon among experts as well (Gigerenzer et al. 2007; Gigerenzer and Gray 2011). For instance, Wegwarth et al. (2012) found in a survey in the United States that most primary care physicians have severe difficulties understanding which statistics are relevant to assessing whether screening saves lives. When examining how general practitioners interpret the results of a diagnostic test (probability of endometrial cancer given a transvaginal ultrasound), Steurer et al. (2002) found that most physicians strongly overestimated the probability of disease given the positive test result.

Another example is probabilistic thinking in the law. Research indicates that—similar to health care providers—jurors, judges, and lawyers often confuse and misinterpret statistical information (Gigerenzer 2002; Kaye and Koehler 1991; Koehler 1996b; Thompson and Schumann 1987). Overall, research on reasoning with probabilities in legal contexts mirrors the difficulties observed in research on probability judgments in general, including a tendency to neglect base rate information but also the opposite, giving too little weight to the evidence (Gigerenzer 2002; Thompson and Schumann 1987).

5.1 Improving Physicians' Diagnostic Reasoning Through Natural Frequencies

Hoffrage and Gigerenzer (1998) investigated the impact of presentation format in the diagnostic reasoning of experienced physicians with different backgrounds (e.g., gynecologists, radiologists, and internists). Physicians were presented with diagnostic inference tasks, such as estimating the probability of breast cancer given a positive mammogram or the probability of colorectal cancer given a positive Hemoccult test. For each problem, the relevant information on base rates, the probability of a positive test result given the disease, and the probability of a positive test given no disease was provided either in terms of probabilities or as natural frequencies. The results showed a dramatic difference between the two presentation formats, increasing from an average of 10 % correct solutions when reasoning with probabilities to an average of 46 % correct solutions when reasoning with natural frequencies. These findings show that experts, too, find it much easier to make probabilistic diagnostic inferences when information is presented in terms of natural frequencies than with probabilities.

In Labarge et al.'s (2003) study with 279 members of the National Academy of Neuropsychology, the physicians' task was to estimate the probability that a patient has dementia, given a positive dementia screening score. Information on the base rate of the disease and the test characteristics were provided in terms of either probabilities or natural frequencies. When information was conveyed through probabilities, only about 9 % of the physicians correctly estimated the posterior probability. By contrast, if the information was provided in a natural frequency format, 63 % correct answers were obtained.

Bramwell et al. (2006) examined the influence of presentation formats on the interpretation of a screening test for Down syndrome. Their findings show that, with probabilities, only 5 % of obstetricians drew correct conclusions, but, when information was presented as natural frequencies, 65 % gave the correct answer. However, their findings also show that other stakeholders in screening (e.g., pregnant women, midwives) had difficulties drawing correct conclusions, even with frequency information. This points to the importance of systematically training health care providers in reasoning with probabilistic information, possibly by using even more intuitive presentation formats such as visual representations of frequency information (see below).

5.2 Probabilistic Thinking in the Law

Advances in forensic science have made the use of DNA analyses a common practice in legal cases, requiring jurors and judges to makes sense of statistical information presented by the prosecution or the defense. In the O. J. Simpson case, for instance, numerous pieces of genetic evidence from the crime scene were introduced during testimony, such as blood matches between samples from the crime scene and

the defendant. Typically, each piece of supposed evidence was presented along with numerical information, such as the probability that the DNA profile of a randomly selected person would match a genetic trace found at the crime scene (so-called "random match probability", see Weir 2007). In one afternoon alone, the prosecution presented the jury with 76 different quantitative estimates (Koehler 1996b).

As in the medical domain, one goal should be to present statistical information in a transparent manner in order to avoid confusion and assist decision makers in making better inferences. This is of particular importance in legal cases, where the prosecutor and the defense may present information in a strategic way to influence the judge or the jury in one way or the other (Gigerenzer 2002; Thompson and Schumann 1987). Lindsey et al. (2003; see also Hoffrage et al. 2000) examined the effects of presenting information on forensic evidence in different presentation formats with a sample of advanced law students and professional jurists. The goal was to compare probabilistic reasoning when numerical information is conveyed through probabilities and when conveyed through natural frequencies. Consider the following example (Lindsey et al. 2003):

> In a country the size of Germany, there are as many as 10 million men who fit the description of the perpetrator. The probability of a randomly selected person having a DNA profile that matches the trace recovered from the crime scene is 0.0001 %. If someone has this DNA profile, it is practically certain that this kind of DNA analysis would show a match. The probability that someone who does not have this DNA profile would match in this type of DNA analysis is 0.001 %. In this case, the DNA profile of the sample from the defendant matches the DNA profile of the trace recovered from the crime scene.

Given this information, the probability that someone has a particular DNA profile if a match was obtained with the trace from the crime scene (i.e., $P(\text{profile} \mid \text{match})$) can be computed according to Bayes' rule (Eq. (1)), yielding a posterior probability of about 9 %. However, given this method of presenting statistical information, only 1 % of law students and 11 % of the jurists were able to derive the correct answer.

Do natural frequencies help understand the numerical information? The same information as above, expressed in terms of natural frequencies, reads as follows:

> In a country the size of Germany, there are as many as 10 million men who fit the description of the perpetrator. Approximately 10 of these men would have a DNA profile that matches the trace recovered from the crime scene. If someone has this DNA profile, it is practically certain that this kind of DNA analysis would show a match. Out of 9,999,990 people who do not have this DNA profile, approximately 100 would be shown to match in this type of DNA analysis. In this case, the DNA profile of the sample from the defendant matches the DNA profile of the trace recovered from the crime scene.

When the statistical information was conveyed this way, significantly more correct answers were obtained (40 % from the law students and 74 % from the professional jurists). This finding is consistent with research on the medical diagnosis task, showing that both legal laypeople (or advanced students, in this case) and experts such as professional jurists can benefit from the use of natural frequencies.

The probative evidence of forensic tests is, of course, not the only factor that plays an important role in legal cases (Koehler 2006). However, presenting evidence so that judges and jurors understand its meaning—and the uncertainties associated with such analyses—is an important prerequisite for making informed decisions.

5.3 Risk Communication: Pictorial Representations

One of the most successful applications of transparent and intuitive presentation formats is risk communication in the health domain. Informed medical decisions—such as whether to participate in a cancer screening program or choosing between alternative treatments—require that both health professionals and patients understand the relevant probabilities.

One example for the importance of understanding quantitative health information concerns the benefits of cancer screening programs, for instance, PSA screening for prostate cancer (Arkes and Gaissmaier 2012). Research shows that both the general public and physicians overestimate the benefits of such programs. For instance, Wegwarth and colleagues (2012) showed that most physicians strongly overestimate the benefits of PSA screening and are led astray by irrelevant statistics. According to a survey with a representative sample of more than 10,000 people from nine countries on the perceived benefits of breast and prostate cancer screening (Gigerenzer et al. 2009), people largely overestimate the benefits of these screening programs. These findings are in stark contrast to the recommendations of health organizations like the U.S. Preventive Services Task Force, who for instance explicitly recommends against PSA-based screening for prostate cancer (Moyer 2012).

What can be done to improve the understanding of the risks and benefits of medical treatments? Using frequencies in either numerical or pictorial form can help people make better, more informed decisions (Akl et al. 2011; Ancker et al. 2006; Edwards et al. 2002; Fagerlin et al. 2005; Gigerenzer and Gray 2011; Gigerenzer et al. 2007; Kurz-Milcke et al. 2008; Lipkus and Hollands 1999). Visualizing statistical information may be especially helpful for people having difficulties in understanding and reasoning with numerical information (Lipkus et al. 2001; Schwartz et al. 1997). Figure 3 gives an example of an *icon array* illustrating the effect of aspirin on the risk of having a stroke or heart attack (Galesic et al. 2009). This iconic representation of simple frequency information visualizes that 8 out of 100 people who do not take aspirin have a heart attack or stroke, as opposed to 7 out of 100 people who do take aspirin. This reduction corresponds to a *relative risk reduction* of about 13 % $[(8 - 7)/8]$. Although communicating the benefits of treatments through relative risk reductions is common practice—in medical journals as well as in patient counseling and the public media—it has been shown to lead to overestimations of treatment benefits (Akl et al. 2011; Bodemer et al. 2012; Covey 2007; Edwards et al. 2001). Galesic and colleagues (2009) showed that iconic representations help both younger and older adults gain a better understanding of relative risk reductions. Icon arrays were particularly helpful for participants with low numeracy skills (Lipkus et al. 2001; Peters 2008; Schwartz et al. 1997).

Galesic and colleagues used icon arrays to visualize simple frequency information (event rates in treatment and control group) and thus aid people in understanding the meaning of relative risk reduction. Another example of an icon array visualizing information on the benefits and harms of PSA screening is shown in Fig. 4 (cf. Arkes and Gaissmaier 2012). It contrasts two groups of people, men who participate in PSA screening versus men who do not. The icon array provides information on

Fig. 3 Icon array used by Galesic et al. (2009) to visualize the effect of taking aspirin on the risk of having a heart attack. The data entail a relative risk reduction of 13 % (from 8 out of 100 to 7 out of 100)

For people with symptoms of arterial heart disease, aspirin can reduce the risk of having a heart attack or stroke by 13%.

Without aspirin   With aspirin

how many men out of 1,000 aged 50 and older are expected to die from prostate cancer in each of the groups, as well as the number of false alarms and number of men diagnosed and treated for prostate cancer unnecessarily. This visual display provides a transparent and intuitive way of communicating the potential benefits and harms of a medical treatment.

Prostate Cancer Early Detection

HARDING CENTER FOR RISK LITERACY

by PSA screening and digital-rectal examination.
Numbers are for men aged 50 years or older, not participating vs. participating in screening for 10 years.

1,000 men without screening: **1,000 men with screening:**

	Without screening	With screening
🔴 Men dying from prostate cancer:	8	8
🔴 Men dying from any cause:	200	200
🔵 Men that were diagnosed and treated for prostate cancer unnecessarily:	–	20
🔵 Men without cancer that got a false alarm and a biopsy:	–	180
⚪ Men that are unharmed and alive:	800	600

Fig. 4 Example of an icon array visualizing the benefits and harms of prostate-specific antigen (PSA) screening for men age 50 and older. The epidemiological data visualized here is taken from Djulbegovic et al. (2010). Copyright 2012 by the Harding Center for Risk Literacy (www.harding-center.com)

However, one should also note that not all graphical aids are equally effective (Ancker et al. 2006). For instance, Brase (2009) compared icon arrays with two types of Venn diagrams, one with and one without individuating information (i.e., dots in the Venn diagram, with each dot representing an individual person). His findings show a consistent advantage of iconic representations over both types of Venn diagrams, pointing to a special role of iconic representations and the importance of choosing the right visual representation.

6 Teaching Representations

How can the findings from cognitive psychology be used to effectively teach probabilistic thinking? We believe that there are two important lessons to be learned from psychological research. First, when information is numerically presented in terms of probabilities, people have great difficulties in making sound inferences. Second, this difficulty can be overcome by conveying information through natural frequencies rather than (conditional) probabilities. Creating an alternative representation of the problem facilitates much better understanding of—and reasoning with—statistical information.

However, whereas natural frequencies strongly improve people's performance, research also indicates that performance is not perfect. For instance, in the Gigerenzer and Hoffrage (1995) study, natural frequencies elicited around 50 % correct responses. This finding is impressive compared to the number of correct responses when reasoning with conditional probabilities, but performance can be further improved by systematically teaching people how to use the power of presentation formats.

Sedlmeier and Gigerenzer (2001; see also Cole and Davidson 1989; Kurz-Milcke and Martignon 2006; Kurzenhäuser and Hoffrage 2002; Sedlmeier 1999) developed a computer-based tutorial program for teaching Bayesian reasoning. The key feature was to teach participants to translate statistical problems into presentation formats that were more readily comprehensible. The study compared rule-based training (i.e., learning how to plug in probabilities into Bayes' rule) with two types of frequency formats: a natural frequency tree (as in Fig. 1 left) and an icon array (Sedlmeier and Gigerenzer used the term frequency grid). A control group without training served as a base line condition. Both short- and long-term training effects were assessed, as was the generalization to new (elemental Bayesian reasoning) problems.

The basic setup of the training procedure was as follows: First, participants were presented with two text problems (the mammography problem and the sepsis problem; see Fig. 5). In the *rule condition*, participants learned how to insert the probabilities stated in the problem in Bayes' rule (Eq. (1)). In the two frequency representation conditions, participants learned how to translate the given probabilities into a natural frequency tree (Fig. 1, left) or an icon array (Fig. 5).

After being guided through the first two problems, participants were presented with eight further problems. For each problem, their task was to solve the problem

100 patients
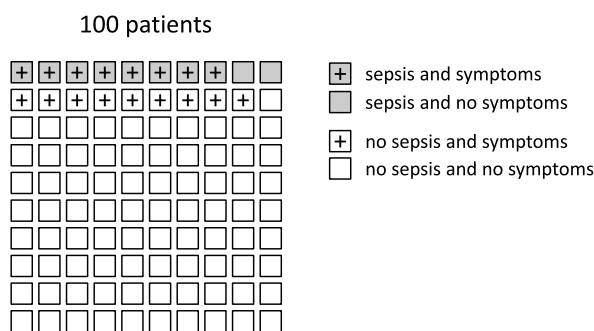
Fig. 5 Icon array similar to the one Sedlmeier and Gigerenzer (2001) used to train participants on translating probabilities into natural frequencies. The task is to infer the posterior probability of sepsis given the presence of certain symptoms, $P(\text{sepsis} \mid \text{symptoms})$. Each square represents one individual. The icon array illustrates the base rate of sepsis in a sample of 100 patients (grey squares) and the likelihood of the symptoms (denoted by a "+") in patients with and without sepsis. The frequencies correspond to the following probabilities: $P(\text{sepsis}) = 0.1$, $P(\text{symptoms} \mid \text{sepsis}) = 0.8$, $P(\text{symptoms} \mid \text{no sepsis}) = 0.1$

(i.e., to compute the posterior probability) by inserting the probabilities into Bayes' rule or by creating a frequency representation (tree vs. icon array, respectively). When participants had difficulties solving these problems, the computer provided help or feedback.

The effectiveness of the training methods was assessed through three post-training sessions (immediate, 1-week follow-up, 5-weeks follow-up). The results showed that teaching participants to translate probabilities into natural frequencies was the most effective method. For instance, prior to training, the median percentage of correct solutions was 0 % (rule condition) and 10 % (natural frequency conditions). When tested immediately after training, the median number of correct solutions was 60 % in the rule condition, 75 % in the icon array condition, and 90 % in the natural frequency tree condition. The most important findings concern the stability of the training effects over time. In the rule-based training, participants' performance was observed to decrease over time; after 5 weeks, the median percentage of correct solutions was reduced to 20 %. For participants who had been trained to use natural frequency formats, by contrast, no such decrease was observed.

7 Conclusions

Thinking and reasoning with statistical information is a challenge for many of us. Many researchers are (or have been) of the opinion that people are severely limited in their capacity of sound probabilistic thinking and fall prey to "cognitive illusions" (Edwards and von Winterfeldt 1986). However, whereas visual illusions may be an unavoidable by-product of the perception system, cognitive illusions are not hard-wired (for a detailed discussion, see Gigerenzer 1996; Kahneman and Tversky 1996). Recent research has revealed insights on how to help people—adults

and children, laypeople and experts—to reason better with probabilistic information about risk and uncertainties. The insights gained from this line of research have also been used to inform applied research, such as how to effectively communicate risks and benefits of medical treatments to health professionals and patients. The most important lesson learned from this research is that information formats matter—in fact, they matter strongly.

7.1 Implications for Mathematics Education

Children, even at a young age, have basic intuitions about probability concepts, such as how the proportion of events relates to the outcomes of simple sampling processes (Brainerd 1981; Fischbein et al. 1970; Girotto and Gonzalez 2008; Yost et al. 1962). One goal of mathematics education should be to foster these intuitions from elementary school on, so that they can later serve as a basis for the acquisition of probability calculus.

A promising road to advance such intuitions is to use playful activities, for example, assembling the so-called "tinker cubes" to illustrate and compare different proportions (Martignon and Krauss 2007, 2009; Kurz-Milcke et al. 2008; see also Martigon 2013). Such intuitive representations can also be used to illustrate the relation between feature distributions and simple sampling processes (Kurz-Milcke and Martignon 2006). Later, when children are taught the basics of probability theory, these intuitions can help pupils develop an understanding of probabilities and simple statistical inferences. Teaching concepts like conditional probabilities, in turn, should use real-world examples and capitalize on the power of natural frequencies and visual aids such as icon arrays and trees. Children and adults alike should be taught representations and not merely the application of rules.

7.2 The Art of Decision Making

Making good decisions requires more than just number crunching. First, it is very important to develop an understanding of the very concepts to which the numbers refer. For instance, research on the perceived benefits of PSA screening shows that physicians are often led astray by statistical evidence that is irrelevant to the question of whether someone should participate in screening (Wegwarth et al. 2012, 2011). Here and in other situations, a qualitative understanding of the concepts behind the numerical information is a prerequisite for informed decisions. Second, in many situations, reliable probability estimates are not available and cannot be relied upon in the decision making process. Such situations are called "decision making under uncertainty" as opposed to "decision making under risk" (Knight 1921/2006) and require more (and other) cognitive tools than probability. When making decisions under uncertainty, simple heuristics are the tools that people use—and should use (Gigerenzer et al. 1999, 2011).

We therefore believe that teaching children the basics of decision making should be an integral part of education. For instance, it is important to understand that there is no unique, always-optimal way of making decisions. Some situations may require a deliberative reasoning process based on statistical evidence, while other situations might require relying on gut feelings (Gigerenzer 2007). One step toward learning the art of decision making is to understand that a toolbox of decision-making strategies exists that can help people deal with a fundamentally uncertain world (Gigerenzer and Gaissmaier 2011; Gigerenzer et al. 2011; Hertwig et al. 2012).

7.3 Final Thoughts

Two centuries ago, few could imagine a society where everyone could read and write. Today, although the ability to reckon with risk is as important as reading and writing, few can imagine that the widespread statistical illiteracy can be vanquished. We believe that it can be done and that it needs to be done. Learning how to understand statistical information and how to use it for sound probabilistic inferences must be an integral part of comprehensive education, to provide both children and adults with the risk literacy needed to make better decisions in a changing and uncertain world.

References

Akl, E. A., Oxman, A. D., Herrin, J., Vist, G. E., Terrenato, I., Sperati, F., Costiniuk, C., Blank, D., & Schünemann, H. (2011). Using alternative statistical formats for presenting risks and risk reductions. *Cochrane Database of Systematic Reviews*, *3*.

Ancker, J. S., Senathirajah, Y., Kikafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: a systematic review. *Journal of the American Medical Informatics Association*, *13*, 608–618.

Arkes, H. R., & Gaissmaier, W. (2012). Psychological research and the prostate-cancer screening controversy. *Psychological Science*, *23*, 547–553.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*, 211–233.

Bastian, L. A., Nanda, K., Hasselblad, V., & Simel, D. L. (1998). Diagnostic efficiency of home pregnancy test kits: a meta-analysis. *Archives of Family Medicine*, *7*, 465–469.

Bodemer, N., Meder, B., & Gigerenzer, G. (2012). Communicating relative risk changes with baseline risk: the effect of presentation format and numeracy. Manuscript under revision.

Bramwell, R., West, H., & Salmon, P. (2006). Health professionals' and service users' interpretation of screening test results: experimental study. *British Medical Journal*, *333*, 284–289.

Brase, G. L. (2009). Pictorial representations and numerical representations in Bayesian reasoning. *Applied Cognitive Psychology*, *23*, 369–381.

Brainerd, C. J. (1981). Working memory and the developmental analysis of probability judgment. *Psychological Review*, *88*, 463–502.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *The New England Journal of Medicine*, *299*, 999–1001.

Christensen-Szalanski, J. J. J., & Beach, L. R. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance*, *29*, 270–278.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology. Human Perception and Performance*, *7*, 928–935.

Cole, W. G., & Davidson, J. E. (1989). Graphic representation can lead to fast and accurate Bayesian reasoning. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, *8*, 227–231.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions of the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.

Covey, J. (2007). A meta-analysis of the effects of presenting treatment benefits in different formats. *Medical Decision Making*, *27*, 638–654.

Daston, L. J. (1988). *Classical probability in the enlightenment*. Princeton: Princeton University Press.

Djulbegovic, M., Beyth, R. J., Neuberger, M. M., Stoffs, T. L., Vieweg, J., Djulbegovic, B., & Dahm, P. (2010). Screening for prostate cancer: systematic review and meta-analysis of randomized controlled trials. *British Medical Journal*, *341*, c4543.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: heuristics and biases* (pp. 249–267). Cambridge: Cambridge University Press.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York: Wiley.

Edwards, W., & von Winterfeldt, D. (1986). On cognitive illusions and their implications. In H. R. Arkes & K. R. Hammond (Eds.), *Judgment and decision making*. Cambridge: Cambridge University Press.

Edwards, A., Elwyn, G., Covey, J., Matthews, E., & Pill, R. (2001). Presenting risk information—a review of the effects of "framing" and other manipulations on patient outcomes. *Journal of Health Communication*, *6*, 61–82.

Edwards, A., Elwyn, G., & Mulley, A. (2002). Explaining risks: turning numerical data into meaningful pictures. *British Medical Journal*, *324*, 827–830.

Evans, J., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, *77*, 197–213.

Fagerlin, A., Wang, C., & Ubel, P. A. (2005). Reducing the influence of anecdotal reasoning on people's health care decisions: is a picture worth a thousand statistics? *Medical Decision Making*, *25*, 398–405.

Fischbein, E., Pampu, I., & Minzat, I. (1970). Comparison of ratios and the chance concept in children. *Child Development*, *41*, 377–389.

Galesic, M., García-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychology*, *28*, 210–216.

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

Gigerenzer, G. (2002). *Calculated risks: how to know when numbers deceive you*. New York: Simon & Schuster.

Gigerenzer, G. (2007). *Gut feelings: the intelligence of the unconscious*. New York: Viking Press.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gigerenzer, G. & Gray, J. A. M. (Eds.) (2011). *Better doctors, better patients, better decisions: envisioning health care 2020*. Cambridge: MIT Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, *102*, 684–704.

Gigerenzer, G., & Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Meilers and McGraw (1999). *Psychological Review*, *106*, 425–430.

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.) (2011). *Heuristics: the foundations of adaptive behavior*. New York: Oxford University Press.

Gigerenzer, G., Mata, J., & Frank, R. (2009). Public knowledge of benefits of breast and prostate cancer screening in Europe. *Journal of the National Cancer Institute*, *101*, 1216–1220.

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

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96.

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form. *Cognition*, *78*, 247–276.

Girotto, V., & Gonzalez, M. (2008). Children's understanding of posterior probability. *Cognition*, *106*, 325–344.

Hammerton, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology*, *101*, 252–254.

Hertwig, R., Hoffrage, U., & the ABC Research Group (2012). *Simple heuristics in a social world*. New York: Oxford University Press.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*, 538–540.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, *84*, 343–352.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science*, *290*, 2261–2262.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J.-P. (1999). Naive probability: a mental model theory of extensional reasoning. *Psychological Review*, *106*, 62–88.

Kahneman, D., & Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, *103*, 582–591.

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

Kaye, D. H., & Koehler, J. J. (1991). Can jurors understand probabilistic evidence? *Journal of the Royal Statistical Society. Series A*, *154*, 75–81.

Kleiter, G. (1994). Natural sampling: rationality without base rates. In G. H. Fischer & D. Lang (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 375–388). New York: Springer.

Knight, F. H. (1921/2006). *Risk, uncertainty and profit*. Boston: Houghton Mifflin.

Koehler, J. J. (1996a). The base rate fallacy reconsidered: descriptive, normative and methodological challenges. *Behavioral and Brain Sciences*, *19*, 1–54.

Koehler, J. J. (1996b). On conveying the probative value of DNA evidence: frequencies, likelihood ratios and error rates. *University of Colorado Law Review*, *67*, 859–886.

Koehler, J. J. (2006). Train our jurors. In G. Gigerenzer & C. Engel (Eds.), *Heuristics and the law* (pp. 303–326). Cambridge: MIT Press.

Kurz-Milcke, E., Gigerenzer, G., & Martignon, L. (2008). Transparency in risk communication: graphical and analog tools. *Annals of the New York Academy of Sciences*, *1128*, 18–28.

Kurz-Milcke, E., & Martignon, L. (2006). Lebendige Urnen und ereignisreiche Bäume: Überlegungen und Versuche zu einer Didaktik der Stochastik in der Grundschule [Lively urns and

eventful trees: developing a didactics of stochastics in elementary school]. *Anregungen zum Stochastikunterricht*, *3*, 182–203.

Kurzenhäuser, S., & Hoffrage, U. (2002). Teaching Bayesian reasoning: an evaluation of a classroom tutorial for medical students. *Medical Teacher*, *24*, 516–521.

Labarge, A. S., McCaffrey, R. J., & Brown, T. A. (2003). Neuropsychologists' abilities to determine the predictive value of diagnostic tests. *Archives of Clinical Neuropsychology*, *18*, 165–175.

Lewis, C., & Keren, G. (1999). On the difficulties underlying Bayesian reasoning: comment on Gigerenzer and Hoffrage. *Psychological Review*, *106*, 411–416.

Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics*, *43*, 147–163.

Lipkus, I. M., & Hollands, J. G. (1999). The visual communication of risk. *Journal of the National Cancer Institute Monographs*, *25*, 149–163.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*, 37–44.

Martigon, L. (2013). Fostering children's probabilistic reasoning and their risk evaluation. In E. J. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking: presenting plural perspectives*. Dordrecht: Springer. This volume.

Martignon, L., & Krauss, S. (2009). Hands on activities with fourth-graders: a tool box of heuristics for decision making and reckoning with risk. *International Electronic Journal for Mathematics Education*, *4*, 117–148.

Martignon, L., & Krauss, S. (2007). Gezinkte Würfel, Magnetplättchen und Tinker Cubes für eine Stochastik zum Anfassen in der Grundschule. *Stochastik in der Schule*, *27*, 21–30.

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*, 119–148.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2009). A rational model of elemental diagnostic inference. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 2176–2181). Austin: Cognitive Science Society.

Moyer, V. A., on behalf of the U.S. Preventive Services Task Force (2012). Screening for prostate cancer: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*, *157*, 120–134.

Multmeier, J. (2012). *Representations facilitate Bayesian reasoning: computational facilitation and ecological design revisited*. Unpublished doctoral dissertation, Free University, Berlin.

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: information acquisition optimizes probability gain. *Psychological Science*, *21*, 960–969.

Peters, E. (2008). Numeracy and the perception and communication of risk. *Annals of the New York Academy of Sciences*, *1128*, 234–267.

Peterson, C. R., & Beach, L. R. (1967). Man as intuitive statistician. *Psychological Bulletin*, *68*, 29–46.

Peterson, C. R., & Miller, A. J. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, *70*, 117–121.

Peterson, C. R., Schneider, R. J., & Miller, A. J. (1965). Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology*, *69*, 522–527.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*, 346–354.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. New York: Norton. (Original work published 1951).

Rouanet, H. (1961). Études de décisions expérimentales et calcul de probabilités [Studies of experimental decision making and the probability calculus]. In *Colloques Internationaux du Centre National de la Recherche Scientifique* (pp. 33–43). Paris: Éditions du Centre National de la Recherche Scientifique.

Schwartz, L., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, *127*, 966–972.

Sedlmeier, P. (1999). *Improving statistical reasoning: theoretical models and practical implications*. Mahwah: Erlbaum.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology. General*, *130*, 380–400.

Simon, H. (1969). *The science of the artificial*. Cambridge: MIT Press.

Steurer, J., Fischer, J. E., Bachmann, L. M., Koller, M., & ter Riet, G. (2002). Communicating accuracy of tests to general practitioners: a controlled study. *British Medical Journal*, *324*, 824–826.

Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials. *Law and Human Behavior*, *11*, 167–187.

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

Wegwarth, O., Schwartz, L. M., Woloshin, S., Gaissmaier, W., & Gigerenzer, G. (2012). Do physicians understand cancer screening statistics? A national survey of primary care physicians. *Annals of Internal Medicine*, *156*, 340–349.

Wegwarth, O., Gaissmaier, W., & Gigerenzer, G. (2011). Deceiving numbers: survival rates and their impact on doctors' risk communication. *Medical Decision Making*, *31*, 386–394.

Weir, B. S. (2007). The rarity of DNA profiles. *Annals of Applied Statistics*, *1*, 358–370.

Yost, P. A., Siegel, A. E., & Andrews, J. M. (1962). Non-verbal probability judgments by young children. *Child Development*, *33*, 769–780.

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: the role of representation in mental computation. *Cognition*, *98*, 287–308.