CHAPTER 18

TECHNOLOGY NEEDS PSYCHOLOGY: HOW NATURAL FREQUENCIES FOSTER INSIGHT IN MEDICAL AND LEGAL EXPERTS

RALPH HERTWIG AND ULRICH HOFFRAGE

Abstract

Modern diagnostic technologies such as cancer screening tests or forensic DNA tests confront users with a basic problem, namely that of inferring the accuracy of a diagnostic test on the basis of statistical information about the test. The required statistics can be presented in a wide variety of ways. While these modes of presentation of statistical information are mathematically equivalent, psychologically they are not. In this chapter, we show that supplementing diagnostic technology with psychological knowledge about how people process frequency information provides us with a very simple—but powerful—method for improving diagnostic inferences.

New technologies often exceed the limits of our imagination. As a result, their advent can evoke unanticipated public responses, ranging from repulsion and hostility to fascination and even mystical attributions. Let us use the debut of the X-ray machine as an illustration. According to an historian of medicine, Joel Howell (1995, pp. 135–137), at first many people were merely fascinated by it. They lined up for one hour sittings to view their own bones. Coin-operated machines let people glimpse the insides of their hands and feet. Wealthy young women had X-ray pictures taken of themselves holding hands with their betrothed. Not everybody, however, responded so enthusiastically, and as we know in hindsight, so riskily. In particular, the power of the X-ray machine to invade privacy met with alarm. Coming close upon the heels of other technological inventions of the period, including the telephone and the photograph, the X-ray was sometimes viewed as evil in its ability to, in the words of one critic cited by Howell (1995, p. 140) 'render privacy a mere tradition of an unscientific past'.

What this mixture of fascination about and hostility toward technological innovation suggests is that new technologies often outstrip our ability to use them properly, let alone to comprehend them completely. If using the technology requires no specific expertise, however, then such lack of comprehension need not be of concern. In fact, many technological tools are designed to be easy to use. Think of the microwave oven or the television: few of us know how either actually works, but that does not mean we cannot use them properly—basically, all we need to know is how to switch them on and off. Sometimes, however, the proper use of a new technological tool or of its end product cannot be reduced to pushing a button but instead requires specific knowledge that is not necessarily part of the human mind's intuitive repertoire. Our argument is that when this is the case, technology needs psychology because the best technology is of little value if people do not comprehend what the results it produces mean.

Some technological tools yield results that need to be interpreted in the context of statistical information. We will show that understanding how laypeople and experts tend to process frequency and probability information can greatly improve people's comprehension of these results. The tools we are concerned with are (1) medical diagnostic tests (e.g. hemoccult test), which play an ever more important role in the diagnosis of specific diseases (e.g. colorectal cancer) and (2) forensic DNA analysis, which in the 1990s revolutionized the criminal investigation process. Note that in the present context the user of the tool is not the lab technician who conducts the actual analysis but, for example, the doctor, patient, judge, or juror. They receive information gleaned from the technology—for instance, a mammography test or a DNA match—and they have to determine what these results could possibly mean.

Medicine: how to improve the use of diagnostic tests

What does a positive medical test result—for instance, a positive mammography test—mean? Multiple studies suggest that physicians often do not properly infer the probability of a disease given a positive test result (Casscells *et al.* 1978). In a seminal article on statistical inferences based on results of mammography tests, David Eddy (1982) reported an informal study in which he provided physicians with information that can be summarized as follows (numbers are rounded): for a woman at age 40 who participates in routine screening, the probability of breast cancer is 0.01. If a woman has breast cancer, the probability is 0.8 that she will have a positive mammogram. If a woman does not have breast cancer, the probability is 0.1 that she will still have a positive mammogram. Now, imagine a randomly drawn woman from this age group with a positive mammogram. What is the probability that she actually has breast cancer?

This probability, also called the *positive predictive value* (PPV) of a test, can be calculated from Bayes' rule. This rule is named after Thomas Bayes (1702–1761), an English dissident minister, to whom the solution of the problem of how to make an inference from data to hypothesis is attributed (Stigler 1983). Equation 1 represents Bayes' rule, applied to the medical context:

$$\mathrm{PPV}=\frac{p(\text{disease})\,p(\text{pos}|\text{disease})}{p(\text{disease})\,p(\text{pos}|\text{disease})+p(\neg\text{disease})\,p(\text{pos}|\neg\text{disease})} \qquad (1)$$

In this equation, p(disease) is the *base rate* (or prevalence) of the disease (0.01 in Eddy's example); p(pos|disease) is the *hit rate* (or sensitivity) of the test, that is, the proportion of positive results among people suffering from the disease (0.8); and $p(\text{pos}|\neg\text{disease})$ is the *false-positive rate* of the test, that is, the proportion of positive results among people not suffering from the disease (0.1). Inserting the statistical information into Bayes' rule results in a positive predictive value for a mammography test of 0.075.

$$\text{PPV} = \frac{(0.01)(0.8)}{(0.01)(0.8)+(0.99)(0.1)} = 0.075$$

Yet most of the physicians in Eddy's study (95 out of 100) estimated the positive predictive value of the test to be between 0.7 and 0.8. That is, their estimates of the probability of breast cancer given a positive mammography test exceeded the correct value by a factor of 10. Eddy (1982) argued that the physicians drew the wrong inference based on mammography because they had confused the hit rate of the test with its positive predictive value. In his view, 'these errors threaten the quality of medical care' (p. 249). Given this and other demonstrations that physicians make errors when interpreting the outcomes of medical diagnostic tests (see also Windeler & Köbberling 1986), what can be done to improve their inferences? As we argue next, supplementing diagnostic technology with psychological knowledge about how people process probability and frequency information provides us with a very simple—but powerful—method for improving diagnostic inferences.

Natural frequencies help in making diagnostic inferences

Studies concluding that physicians (Berwick *et al.* 1981; Politser 1984) and laypeople (see Koehler 1996) make poor diagnostic inferences based on statistical information have typically presented information in the form of probabilities and percentages. From a mathematical viewpoint, it is irrelevant whether statistical information is presented in probabilities, percentages, absolute frequencies, or some other form because these different representations can be mapped onto one another in a one-to-one fashion. From a psychological viewpoint, however, the representation of information matters. Although different representations of statistical information are equivalent mathematically, psychologically they are not. This observation is central to our argument. In particular, we argue that a specific class of representations that we call *natural frequencies* (Hoffrage & Gigerenzer 1998) helps experts to make inferences the Bayesian way.

We now illustrate the difference between probabilities and natural frequencies using the diagnostic problem of inferring the presence of colorectal cancer (C) from a positive result of the hemoccult test (T), a standard diagnostic test. Let us assume that, in terms of *probabilities*, the base rate for colorectal cancer $p(\text{cancer})$ is 0.003; the test's hit rate, $p(\text{pos}|\text{cancer})$, is 0.5; and the false-positive rate, $p(\text{pos}|\neg\text{cancer})$, is 0.03. Armed with this information, people are then typically asked, 'What is the probability that a randomly drawn person who tested positive actually has colorectal cancer?'

In *natural frequencies*, in contrast, the same information would read 'Thirty out of every 10,000 people have colorectal cancer. Of these 30 people with colorectal cancer, 15 will have a positive hemoccult test. Of the remaining 9,970 people *without* colorectal cancer, 300 will still have a positive hemoccult test.' The question is, 'How many of the people who tested positive actually have colorectal cancer?'

What exactly are natural frequencies? They are absolute frequencies of events as directly experienced and they have *not been normalized with respect to the base rates* of the disease and its absence (Gigerenzer & Hoffrage 1995, 1999). For example, imagine an old, experienced physician in an illiterate society. She has no books or statistical surveys and therefore must rely solely on her direct experience. Her people, for instance, may have been afflicted by a previously unknown and severe disease. Fortunately, the physician has discovered a symptom that signals the disease, although not with certainty. In her lifetime, she has seen a large group of patients, few of whom had the disease. Of those who had the disease, some showed the symptoms; of those who were not afflicted, some also showed the symptoms. Thus, on the basis of her experience, the physician acquired representative information about the structure of her environment by sequentially encountering (randomly drawn) instances in the population. This is what we call natural sampling. The outcome of natural sampling is natural frequencies. Natural frequencies are not to be confused with probabilities, percentages, relative frequencies, or other representations where the underlying natural frequencies have been normalized with respect to these base rates.

Why and how should natural frequencies facilitate diagnostic inferences? There are two related explanations (for alternative views see Fiedler *et al.* 2000; Macchi 2000; for a discussion of those alternative views see Hoffrage *et al.* in press). The first is computational. Bayesian computations are simpler when the information is represented in natural frequencies than in probabilities, percentages, or relative frequencies (Christensen-Szalanski & Bushyhead 1981; Kleiter 1994). Consider the calculations when the information concerning colorectal cancer is represented in probabilities. A cognitive algorithm to compute the PPV of the hemoccult test based on probabilities amounts to Equation 1:

$$\mathrm{PPV} = \frac{p(\text{cancer})\, p(\text{pos}|\text{cancer})}{p(\text{cancer})\, p(\text{pos}|\text{cancer}) + p(\neg\text{cancer})\, p(\text{pos}|\neg\text{cancer})}$$

$$= \frac{(0.003)(0.5)}{(0.003)(0.5)+(0.997)(0.03)} = 0.048$$

Now compare these computations with those necessary when the same information is presented in natural frequencies. Because natural frequencies do not require figuring in base rates, the computations are much simpler—that is, fewer operations (multiplication, addition, or division) need to be performed, and the operations can be performed on natural numbers rather than fractions. Now the algorithm amounts to Equation 2:

$$\mathrm{PPV} = \frac{\text{pos\&cancer}}{\text{pos\&cancer} + \text{pos\&}\neg\text{cancer}} = \frac{15}{15+300} = 0.048 \qquad (2)$$

Equation 2 is Bayes' rule for natural frequencies, where pos&cancer is the number of cases with cancer and a positive test, and pos&¬cancer is the number of cases without cancer but with a positive test.

The second argument as to why natural frequencies facilitate diagnostic inferences complements the first. It suggests that people's cognitive algorithms are designed to make inferences from natural frequencies rather than from probabilities and percentages. This argument is based on the observation that for most of their existence, humans and animals have had to make inferences based on information encoded sequentially through their direct experience. Natural frequencies are the result of this process. Mathematical probability, in contrast, emerged only in the mid-seventeenth century (Daston 1988), and percentages seem to have become common representations only in the aftermath of the French revolution—when the metric system was adopted—mainly for calculating taxes and interest, and only very recently for expressing risk and uncertainty (Gigerenzer *et al.* 1989). Thus one might argue that minds have evolved to deal with natural frequencies rather than with probabilities. This argument is consistent with developmental studies indicating the primacy of reasoning about discrete numbers and counts over fractions, and with studies of adult humans and animals showing that they can monitor frequency information in their natural environments in fairly accurate and automatic ways (e.g. Gallistel & Gelman 1992; Jonides & Jones 1992; Real 1991; Sedlmeier *et al.* 1998).

Do natural frequencies improve physicians' statistical reasoning?

Can medical experts' use of diagnostic technologies be improved if they reason in terms of natural frequencies rather than probabilities? The following study with experienced physicians who treat real patients provides an answer. Hoffrage and Gigerenzer (1998) asked 51 physicians from Munich and Düsseldorf to participate. Three physicians did not give numerical responses to the diagnostic tasks they were asked to complete, either because they considered statistical information to be meaningless for medical diagnosis or because they said that they were unable to think with numbers. The remaining 48 physicians had practised medicine for an average of 14 years, had a mean age of 42 years, and worked in university hospitals, in private or public hospitals, or in private practice. The sample included internists, gynaecologists, dermatologists, and radiologists, among other specialists. The physicians' status ranged from directors of clinics to beginning physicians.

Each physician was given four diagnostic tasks. In two of them, the information was presented in probabilities, whereas in the other two it was presented in natural frequencies (task order was systematically varied). The four diagnostic tasks were to infer the presence of (1) colorectal cancer from a positive hemoccult test (see above), (2) breast cancer from a positive mammogram, (3) phenylketonuria from a positive Guthrie test, and (4) ankylosing spondylitis (Bekhterev's disease) from a positive HL-Antigen-B27 (HLA-B27) test (for the texts, see http://www-abc.mpib-berlin.mpg.de/users/hoffrage/papers/4tasks.html).

Each physician gave an estimate for each of the four diagnostic problems. Thus, the study yielded a total of 48 estimates for each problem and 24 estimates for each format of each problem. When a physician's estimate was within five percentage points (or the equivalent in frequencies) of the Bayesian answer, and additional information (from physicians' notes and/or an interview) indicated that the estimate was arrived at by Bayesian reasoning (or a shortcut thereof; see Gigerenzer & Hoffrage 1995) rather than by guessing or other means, then the response was classified as a Bayesian inference.

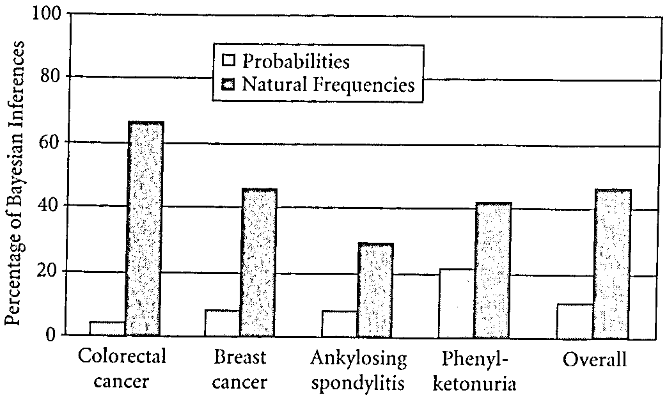

Figure 18.1 Physicians' percentage of Bayesian inferences in the probability and natural frequency versions of four diagnostic tasks (Hoffrage & Gigerenzer 1998).

Figure 18.1 shows that for each diagnostic problem, the physicians reasoned the Bayesian way more often when the information was communicated in natural frequencies than in probabilities. The size of this effect varied between problems, but even in the problem showing the smallest effect (phenylketonuria), the proportion of Bayesian answers was twice as large in the natural frequency than in the probability version. For the two cancer problems, natural frequencies increased Bayesian inferences by more than a factor of five. Across all problems, the physicians gave the Bayesian answer when provided with probabilities in only 10% of the cases; when they were provided with natural frequencies, this value increased to 46%.

As these results demonstrate, natural frequencies are a powerful tool for improving doctors' statistical inferences. Using natural frequencies also paid off in terms of time efficiency and an increased sense of self-efficacy. Given probabilities, the physicians in the study spent an average of 25% more time solving the diagnostic problems than they did when given natural frequencies. Comments made by the physicians revealed that they were more often nervous, tense, and uncertain about solving the tasks when they were working with probabilities than with natural frequencies. In addition, their spontaneous remarks revealed that they were less sceptical about the relevance of statistical information when it was expressed in natural frequencies. The physicians were aware of their better and faster performance with natural frequencies, as illustrated by comments such as the following: 'Now it's different. It's quite easy to imagine. There is a frequency—that's more visual' and 'a first grader could do this!'

Different expressions of statistical information affect how people process and understand information. Are the benefits of representing statistical information in terms of natural frequencies specific to medical inferences or do they generalize beyond the medical domain? We believe they do. A particularly important professional context—in which statistical inferences can literally determine whether a person lives or dies—is that of forensic DNA analysis. Since the early 1990s, this technology has dramatically altered the criminal investigation process in cases where biological evidence (e.g. semen, blood, saliva, hair) is found at the crime scene. Forensic evidence unearthed by this technology is currently challenging the practice of capital punishment in the United States. We next explore the nature of this technology-driven challenge and then ask how natural frequencies can improve statistical inferences drawn from DNA test results.

Forensic DNA analysis and the death penalty

In recent years, the US criminal justice system has applied its most severe penalty—the taking of the convict's life—ever more frequently. In 1999, ninety-eight prisoners were executed for capital crimes in the United States, more than in any other years since 1951 (*The Economist*, 10 June 2000). The increased rate of executions in the United States is particularly striking at a time when the US capital crime rate has been dropping and when, according to Amnesty International, over half of the countries in the world have abolished the death penalty in law or practice. Among the big democracies, only the United States, India, and Japan still put prisoners to death.

The increase in executions also contrasts with an important shift in public opinion documented in a Gallup survey conducted in February 2000. Though a majority of Americans still support the death penalty, the public support has been gradually decreasing since its high point in 1994 and at 66% has reached its lowest level since 1981.[1] Coinciding with the waning support for the death penalty is the growing belief—expressed by 91% of all respondents (i.e. even many of those who favour the death penalty)—that innocent people are at least occasionally wrongly sentenced to death. What fuels this belief in the fallibility of the criminal justice system, according to the Gallup researchers, is the advent of modern DNA technology (which was not used in US courts until 1989; see Wells *et al.* 2000)—specifically the fact that DNA testing has produced new evidence suggesting that innocent people were sentenced to death in American courts.

The public debate concerning the sentencing of innocent people to death has to a large extent been fueled by the *Innocence Project*. Founded in 1992 by two prominent New York criminal defence lawyers, Barry Scheck and Peter Neufeld, this organization at the Yeshiva University's Benjamin N. Cardozo School of Law provides *pro bono* legal assistance to people who challenge their convictions by using DNA testing. In the almost 10 years of its existence, the Innocence Project has represented or assisted in 36 of the 63 cases where convictions have been reversed or overturned in the United States over this period. In a recent book, Scheck, Neufeld, and *The New York Daily News* columnist Jim Dwyer (2000) recount the harrowing stories of men who were wrongfully convicted of crimes and, after many years and much resistance from prosecutors and judges, exonerated by DNA test results. In most of these cases, the charge was rape and the penalty a prison term, though some who were wrongfully convicted of murder were sentenced to death.

The Innocence Project and similar initiatives throughout the United States have stirred up a political debate about the death penalty (see Lifton & Mitchell 2000). In response to the mounting evidence of serious flaws in the capital punishment system, the governor of Illinois, for instance, declared a moratorium on executions in that state. And, for the first time in American legal history, a judge (at the Houston County Superior Court)

[1] For the detailed results of this survey, see http://www.gallup.com/poll/releases/pr000224.asp.

recently authorized DNA analysis of evidence from a death penalty case in which the convicted man has already been put to death. Prior to this order, advocates of the death penalty could argue that there exists no incontrovertible proof that an innocent person was ever wrongfully executed, as did Edwin Meese, Attorney General in the Reagan Administration: 'If a person is innocent of crime, then he is not a suspect' (quoted in Scheck *et al.* 2000, p. xi). The judge's order to analyse the DNA of the executed man will soon determine whether or not this argument is valid and can still be made in the future.

DNA tests—a gold standard for truth?

Wrongful convictions have been exposed in the past, but cases in which DNA analysis led to exoneration after conviction are particularly impressive, 'perhaps, because they all used a single, definitive technology to establish innocence' (Wells *et al.* 2000, p. 589). In Scheck *et al.*'s (2000) view:

> DNA testing is to justice what the telescope is for the stars: not a lesson in biochemistry, not a display of the wonders of magnifying optical glass, but a way to see things as they really are. It is a revelation machine. (p. xv)

Given that Scheck and Neufeld have witnessed the exonerating power of DNA evidence first hand, their enthusiastic appraisal of DNA technology is more than understandable.[2] But it is nevertheless important to realize that, like medical testing, DNA testing is based on statistical information and does not necessarily produce incontrovertible proof. For this reason, insight into the aspects of DNA testing that call for statistical inference is crucial—not least because the belief that it is 'a gold standard for truth telling' (Scheck *et al.* 2000, p. 122) could itself become the source of wrongful convictions. In each of Scheck *et al.*'s exoneration cases, the lack of a match between a convict's DNA profile and traces found at the crime scene excluded him as the source of the trace. But what if a trace had matched? Would that have provided ironclad proof that the convict was the source? As with the outcome of a medical diagnostic test, interpreting the outcome of a DNA analysis requires statistical reasoning, and the statistical reasoning of judges, jurors, and possibly even DNA experts may be facilitated by natural frequencies.

How natural frequencies can help in interpreting DNA evidence

Although it is unlikely that a criminal suspect would coincidentally share a DNA profile with a piece of incriminating evidence, just how unlikely that coincidence is

[2] Scheck and Neufeld's unconditional support for DNA evidence may come as a surprise given the role they played in the O.J. Simpson defence team. There, they succeeded in thoroughly discrediting the DNA evidence presented by the prosecution. How do their past criticism of DNA testing and their present enthusiasm for it go together? In fact, Scheck and Neufeld's position on DNA technology actually has not changed, as a closer look at the Simpson trial transcripts (available on the Internet) shows. In his closing arguments, Scheck did not question the DNA technology per se—'DNA is a sophisticated technology. It is a wonderful technology'—but the evidence analysed in that case. Specifically, he argued that the DNA evidence against Simpson was mishandled and even fabricated.

depends on the frequency of a specific combination of genetic features in a specified reference class. This reference class may be a racial group, or an artificial probability space created by multiplying together the frequencies with which the individual genetic features of the profile appear in a population. The statistic usually reported at trial is the frequency with which the specific combination of genotypic features occurs in a specified reference class (e.g. an expert witness in the O.J. Simpson case stated that Simpon's DNA profile, which matched that of droplets of blood leading from the bodies, occurs in only 1 in 170 million people).[3]

This frequency may be interpreted as the chance that someone selected at random would have the profile in question. Unfortunately, this statistic seems to be widely misinterpreted by judges, jurors, and even DNA experts themselves. The low frequency of a DNA profile in some populations is sometimes misinterpreted, for example, as the likelihood that an accused person is innocent. While the estimated frequency of a DNA profile might be one in 5 billion, one DNA expert testifying in a US court misinterpreted this figure as 'a one in 5 billion chance that anybody else could have committed the crime' (see Koehler 1993, p. 32, for this and many other examples), and in Germany, the President of the *Deutschen Gesellschaft für Rechtsmedizin* claimed that a DNA match identifies a perpetrator with '100% certainty' (for this and other examples from the German legal system, see Krauss & Hertwig 2000). Other DNA experts have misinterpreted the profile frequency as the probability that the DNA evidence came from anyone other than the defendant, leading judges likewise to misuse it in their opinions as, for example, 'the probability of someone else leaving' the genetic trace.

Even if judges, jurors, and DNA experts could avoid such misinterpretations, the estimated frequency of a specific DNA profile would still be misleading for yet another important reason: it ignores the chance of a laboratory error. Despite occasional expert testimony declaring that laboratory errors are impossible, such laboratory errors do occur—with a frequency several orders of magnitude larger than the chance of a coincidental match (e.g. Koehler *et al.* 1995).[4] To see how laboratory errors affect the

[3] Much controversy about the use of DNA fingerprinting has centred on two questions: (1) what is the frequency of a particular type of DNA marker in the appropriate population, and (2) how should the frequencies of individual markers be combined to calculate the probability of a specific person's DNA profile (e.g. Lander 1989; Lander & Budowle 1994)? Each question hinges on the other. For instance, using the 'product rule' to combine the frequencies of the individual markers requires assuming that the individual alleles at different loci can be treated as statistically independent. This assumption has been hotly disputed (e.g. Lewontin & Hartl 1991). If a population (e.g. total US population) is made up of subpopulations with different gene frequencies (e.g. people of Italian, German, Vietnamese, African decent), then independence cannot be assumed. In response to this argument, a 'ceiling principle' has been proposed according to which the probability of a DNA profile is estimated by combining the largest known allele frequencies from a wide variety of populations (in other words, the figures are chosen from the range of probabilities that are most favourable to the accused). According to Kaye (1997), however, the product rule has recently experienced a comeback, at least in situations in which the class of plausible suspects is as broad as a racial group (see also NRC 1996).

[4] In analysing DNA evidence, technical and human errors can occur. On the technical side, enzyme failures, abnormal salt concentrations, and mischievous dirt spots can produce

calculations, let us look at the necessary Bayesian computations. We first provide the information in probabilities and then in natural frequencies.

In terms of probabilities, the information in this hypothetical case is as follows: first, the base rate of the DNA profile is 0.00001. Second, if someone has that DNA profile, a DNA analysis would reliably show it to match any samples that share it (i.e. $p(\text{match}|\text{profile})=1.0$). Third, if the chances of a false-positive laboratory error are as high as sometimes found, the probability of a false-positive match, $p(\text{match}|\neg\text{profile})$, could be 0.003. To calculate the probability of a person having a particular DNA profile given there is a match with the incriminating evidence, Bayes' rule is required:

$$p(\text{profile}|\text{match})=\frac{p(\text{profile})p(\text{match}|\text{profile})}{p(\text{profile})\,p(\text{match}|\text{profile})+p(\neg\text{profile})\,p(\text{match}|\neg\text{profile})}$$

Inserting the statistical information into Bayes' rule results in a probability that the person who matches actually has the profile of 0.003:

$$p(\text{profile}|\text{match}) = \frac{(0.00001)(1.0)}{(0.00001)(1.0) + (0.99999)(0.003)} = 0.003.$$

These computations are relatively complex, and, as shown earlier, they can be simplified when the information is presented in natural frequencies. In natural frequencies, the same information would be expressed as follows: first, in a population of 10 million, a particular DNA profile occurs with a frequency of 10 in a million. Thus, one might expect approximately 100 people in this population to have the DNA profile. Second, if someone has this profile, a DNA analysis would show it to match any samples that share it. Third, owing to false-positive laboratory errors, however, there could be up to 30,000 people in the population who do not have the same DNA profile but who would nonetheless be found to match in the DNA analysis.

To compute the probability of a person's having a particular DNA profile *given* a match, one requires merely a count of the people who actually have the profile out of all the people who match. The calculations amount to solving Equation 2 (here adapted to the DNA context):

$$p(\text{profile}|\text{match})=\frac{\text{match\& profile}}{\text{match\&profile}+\text{match\&}\neg\text{profile}}=\frac{100}{100+30{,}000}=0.003$$

The confusion over statistical evidence reviewed earlier suggests that judges, jurors, and sometimes even DNA experts do not spontaneously understand evidence presented

misleading DNA banding patterns. In terms of human errors, inadvertent switching, mixing, or cross-contamination of samples may lead to false positive errors. The likelihood of these and other errors are estimated on the basis of blind proficiency tests (described in more detail in Koehler 1993, pp. 24–25).

in terms of probabilities. Because Bayesian calculations are simpler when numbers are expressed as natural frequencies, these expressions may yield insight into the uncertainties of forensic scientific analyses with laboratory errors, but also with the choice of the reference class. Ultimately, choosing different ways to express the evidence could influence decisions about guilt and innocence.

Do natural frequencies improve statistical reasoning in the legal context?

Can legal experts profit from natural frequencies in making inferences, just as medical experts do? In a study conducted at the Free University in Berlin, we (Lindsey *et al.* in press; Hoffrage *et al.* 2000) asked 27 professionals who would soon qualify as judges ('jurists') and 127 advanced law students to evaluate two criminal court case files involving rape. In both cases, a DNA match was reported between a DNA sample from the defendant and one recovered from the victim. Aside from this evidence there was little reason to suspect that the defendant was the perpetrator. Expert testimony reported the frequency of the recovered DNA profile as 1 in 1,000,000 and then stated that it was practically certain that the analysis would show a match for a person who indeed had the DNA profile (i.e. the test's hit rate = 100%). The expert also reported the rates of technical and human errors that would lead to false-positive results.

The expert stated all the statistics as either probabilities or frequencies (see Appendix). Based upon these statistics all participants had to estimate two probabilities—that of having a particular DNA profile *given* a DNA match and that of being the source of the evidence *given* a DNA match. Immediately after their estimates, the participants rendered a verdict for the case: guilty or not guilty.[5] After reading one case file in one format, each participant was given a second case file with expert testimony in the other format. They then answered the same questions as before.

Similar to physicians' inferences, the estimates of the legal decision makers are strongly affected by how the statistical evidence was presented. Figure 18.2 shows the percentage of Bayesian inferences as a function of information format. We found a similar pattern across all estimates participants were asked to produce. Consider, for example, participants' estimates of the probability that the defendant was actually the source of the trace. When the statistics were expressed as probabilities, only 13% of the professionals and fewer than 1% of the law students made the correct inference. But when the

[5] The two probability judgments and the verdict correspond to the stages in the chain of inferences that arise when DNA evidence is presented in court. From a reported match, one may want to infer (1) the probability that the person for whom the match is reported actually has the DNA profile, (2) the probability that this person is the source of the trace recovered from the crime scene, and finally (3) the probability that the person is guilty. It is important to note that the first two probabilities are *not* sufficient to allow an inference of the probability of guilt. To see why, imagine one knows for certain that a particular person is the source of a DNA sample recovered from a crime scene. Thus, this probability (source given match) would equal 1. However, it is still possible that the person left the trace innocently either before or after the crime was committed or that someone else planted it there.

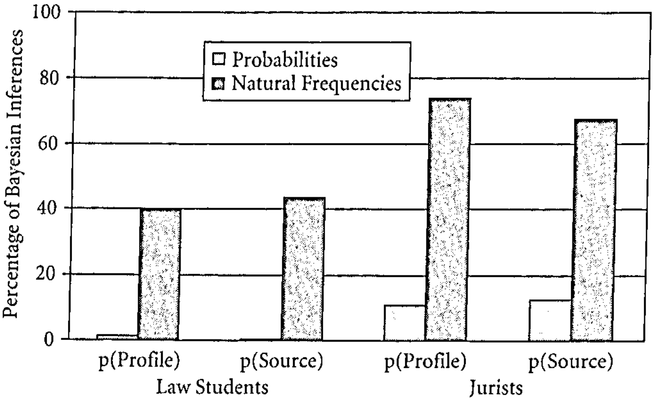

Figure 18.2 Law students' and jurists' percentage of Bayesian inferences in the probability and natural frequency versions (Lindsey *et al.* in press).

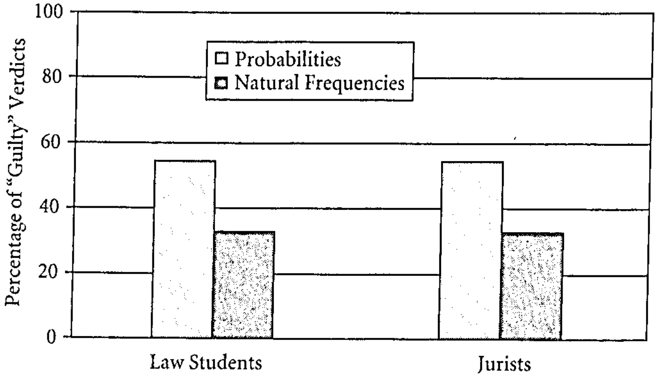

Figure 18.3 Percentage of 'guilty' verdicts of law students and jurists as a function of the representation of information (probability vs natural frequency).

identical statistics were stated as natural frequencies, 68% and 44% of these same participants made the correct inference.[6]

Participants' statistical reasoning also had a clear and important effect on judicial decision making. The mathematically identical statistical evidence led to a higher conviction rate in the probability format than in the natural frequency format. As Fig. 18.3 shows, in both participant samples the proportion of guilty verdicts was substantially higher in the probability format than in the frequency format (13 and 22 percentage points, respectively). Why does the probability format produce more guilty verdicts? There appears to be a simple answer. The estimates that participants calculated from the probability information far exceeded those computed from the frequency information. For instance, jurists, on average, estimated the probability of having the DNA profile in question given a DNA match to be 0.63. In contrast, the average estimated probability in the natural frequency format was 0.05. Thus, it is not surprising that the larger

[6] Based on the statistics reported in the Appendix, the probability that a person who is found to match in a DNA analysis (with the evidence from the crime scene) actually has the DNA profile in question is .09 (i.e. 10/110); the probability that a person is the source of the trace recovered from the crime scene given that he is found to match in the DNA analysis is 0.009 (i.e. 1/110).

estimates that were made based on the probability information—when taken as the source probability—led to a higher proportion of guilty verdicts.

Why did participants make higher estimates when given statistical evidence in probabilities? From participants' written explanations of how they derived their estimates we were able to identify two major non-Bayesian algorithms that participants used in the probability format. In the sample of jurists, for instance, about two-fifths (38%) of *all* responses (not just of those responses where a cognitive algorithm could be identified) were produced by *likelihood subtraction* and the *hit rate minus base rate* algorithm. Likelihood subtraction (see also Gigerenzer & Hoffrage 1995), which involves computing the difference $p(\text{match}|\text{profile}) - p(\text{match}|\neg\text{profile})$, makes no use of base rate information (here the base rate of the profile). In a context in which hit rate is very high, as in the case of forensic DNA analysis, and the false-positive and base rate are relatively low, both non-Bayesian algorithms will thus generate erroneously high probability estimates.

Where to go from here

Although uncertainty is deeply entrenched in many legal and medical decisions, the means to reckon with uncertainty have not necessarily been well understood or appreciated by professionals in either discipline. Take one of the first applications of probability theory to legal evidence in history as an example. The famous *Howland Will Case* figured the female heir to one of the greatest fortunes in the United States, Hetty Robinson, fighting to gain control of every last penny. The trial, which took place shortly after the end of the Civil War, turned into a protracted battle over a single piece of evidence: the signature of Hetty Robinson's aunt on a document that effectively left most of her property to her niece (for a description of the trial and its circumstances, see *The New Yorker*, 23 & 30 April 2001, pp. 62–70). The opposite side claimed that the aunt's signature had been forged. An impressive number of expert witnesses was enlisted to litigate the matter. Among them were Benjamin Peirce, Harvard professor, and his son, Charles Sanders Peirce (who later became a famous logician and philosopher).

To determine the validity of the signature in dispute, the Peirces identified thirty places in the aunt's verified signature where she had made a downstroke with her pen (thus forming a letter). When they superimposed the disputed signature on the verified signature, all thirty downstrokes started at exactly the same point on each letter. How likely is that to happen by chance? To estimate this likelihood they analysed previous signatures of Robinson's aunt and found that, on average, one out of five downstroke positions overlapped. With this base rate, they argued that the chance that Robinson's aunt could have unintentionally produced two signatures in which all thirty downstrokes overlapped, was 1 in 5^{30}—'so vast an improbability, is practically an impossibility . . . . It is utterly repugnant to sound reason to attribute this coincidence to any cause but design' (Benjamin Peirce in his deposition, as cited in the *The New Yorker*, 23 & 30 April 2001, p. 69). The time, however, was not ripe for such reasoning. Both the lawyers for Hetty Robinson and the public treated this argument as mathematical voodoo; it was ridiculed, and some even felt that it transgressed a boundary by applying the laws of probability to 'elements of will and desire unfit... for judgment by such laws' (*The New Yorker*, 23 & 30 April 2001, p. 69).

Are those reactions just reminders of a distant past in which both the legal profession and the public were not educated to think about uncertainties? Unfortunately, there is plenty of evidence that suggests that the justice system has not yet overcome the illusion of certainty. Take eyewitness identification as an example. Eyewitness identification can be highly persuasive to jurors, although it is the major source of wrongful convictions (Borchard 1932; Rattner 1988; Wells *et al.* 2000). Despite this knowledge, the criminal justice system, and in particular the prosecutors (see Wells *et al.* 2000), seem to be utterly reluctant to adopt new (and empirically validated) procedures in collecting eyewitness testimony, which were designed to increase the reliability of eyewitness identification (e.g. sequential lineups; see Wells *et al.* 2000).

Why do legal and medical decision-makers seem to have difficulties reckoning with uncertainties? The problem starts in their training. Medical schools teach statistics, but their focus is on methods of data analysis such as significance testing. But even if they are taught statistical procedures needed for risk assessment, students are typically instructed to mechanically insert probabilities into mathematical formulas such as Bayes' rule. In law, the case for statistical reasoning seems to be even worse. With few exceptions, law schools do not teach students how to reason on the basis of uncertain evidence—although virtually all evidence is uncertain. If, then, statistical illiteracy has its roots in training or in the lack thereof, what can be done? We suggest that the endeavour to bring about statistical literacy is more likely to succeed if training of statistical reasoning is based upon information representations that are suited to the human mind.

It is noteworthy that the beneficial effects of natural frequencies on statistical reasoning in the studies reported above occurred *without* any training or instruction. Naturally, this raises the hope that systematic training in the use of natural frequencies may improve people's ability to reason statistically even more dramatically. The key is to teach representations rather than rules; that is, to solve problems—such as the medical and legal ones described here—by translating probabilities into natural frequencies. In fact, Sedlmeier and Gigerenzer (2001) showed that such 'representation training' can make an enormous difference. In contrast to a traditional 'rule training', their two-hour representation training was both much more successful in improving people's performance in the short run and in keeping people from forgetting how to solve such problems in the long run. Thus these results suggest that the teaching of representations—in high schools, colleges, and universities—can be an important pedagogical tool to faster, more reliable, and more comprehensibly attainable statistical literacy.

Being able to reason statistically is important not only for professionals but also for their clientele, that is, us. During consultation with their patients, for instance, in the United States doctors are increasingly more likely to say 'I can't tell you what to do'. According to George J. Annas, the chairman of the health law department at Boston University's School of Public Health, 'many doctors are comfortable now saying, "It's not me, it's you, and you are the one who has to decide"' (quoted in *The New York Times*, 25 June 2000, Section 15, p. 1). Two factors seem to be driving this transfer of decision-making responsibility from the doctor to the patient. According to medical professors interviewed in a recent article in *The New York Times* (25 June 2000, Section 15, pp. 1, 10), one is the fear of lawsuits: 'Doctors, after all, can be sued over whatever decision they make. But if the patient makes the decision, who is to blame?'. The second factor is the burgeoning variety and complexity of technological tools in medicine: 'In a world where medical technology is getting all the more powered, and often accompanied by risks, nobody can decide for you'.

Radiation-based diagnostic procedures are a case in point. According to Horst Kuni (*Süddeutsche Zeitung*, 3 August 2000, p. B-2), professor of nuclear medicine at the University of Marbug, in Germany up to 50,000 people (!) per year fall ill with cancer because of radiation-based medical examinations of sometimes questionable utility (such as mammography, see Gotzsche & Olsen 2000). What this figure makes abundantly clear is that any exposure to radiation carries risk. This is a fact that neither the public nor scientific experts recognized when they celebrated the advent of the new technology in the early decades of the twentieth century (Howell 1995), but it is well known today. Thus, patients and health care providers must decide on a case-by-case basis whether the information gleaned from a medical diagnostic procedure using radiation justifies its use. Needless to say, it is therefore essential that patients have a proper understanding of the available statistical information (e.g. the positive predictive value of a test). We suggest such an understanding is more likely to be achieved if doctors communicate statistical information in terms of natural frequencies instead of probabilities.

Conclusion

Increasingly, modern technologies are shaping many aspects of our lives. Yet these rapid technological developments have not always delivered their intended benefits—often because of the limited understanding of how the results that the technologies produce ought to be communicated. It is here that technology needs psychology. For instance, to improve doctors' and patients' interpretation of a positive mammogram, we need to understand how the way this result is communicated relates to, and interacts with, the way the human mind works. We showed that psychological research on how people process frequencies allows us to improve the reasoning of those who need to make statistical inferences from the results of diagnostic technologies. Although representing those results one way or another may not make much of a difference for a mathematician, it can make one for a juror and a physician, and ultimately for a defendant and a patient.

Note

We are grateful to Tilmann Betsch, Valerie M. Chase, Steffi Kurzenhäuser, Peter Sedlmeier, and an anonymous reviewer for many helpful comments, and to the Deutsche Forschungsgemeinschaft for its financial support of the first author (Forschungstipendium He 2768/6–1) and the second author (Ho 1847/1).

Correspondence concerning this chapter should be addressed to Ralph Hertwig, Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, Lentzeallee 94, 14195 Berlin, Germany. Electronic mail may be sent to hertwig@mpib-berlin.mpg.de.

References

Berwick, D. M., Fineberg, H. V. & Weinstein, M. C. (1981). When doctors meet numbers. *American Journal of Medicine, 71*:991–998.

Borchard, E. (1932). *Convicting the innocent: Errors of criminal justice.* New Haven: Yale University Press.

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

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.

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

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

Fiedler, K, Brinkmann, B., Betsch, T. & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General, 129*:399–418.

Gallistel, C. R. & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition, 44*:43–74.

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 & Keren and Mellers & McGraw. *Psychological Review, 104*:425–430.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life.* Cambridge, UK: Cambridge University Press.

Gotzsche, P. C. & Olsen, O. (2000). Is screening for breast cancer with mammography justifiable? *Lancet, 355*:129–134.

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. (in press). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition.*

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

Howell, J. D. (1995). *Technology in the hospital: Transforming patient care in the early twentieth century.* Baltimore: Johns Hopkins University Press.

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

Kaye, D. H. (1997). DNA identification in criminal cases: Lingering and emerging evidentiary issues. In *Proceedings of the Seventh International Symposium on Human Identification* (pp. 12–25). Madison: Promega Corp.

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

Koehler, J. J. (1993). Error and exaggeration in the presentation of DNA evidence. *Jurimetrics Journal, 34*:21–39.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19*:1–53.

Koehler, J. J., Chia, A. & Lindsey, S. (1995). The random match probability (RMP) in DNA evidence. *Jurimetrics Journal, 35*:201–219.

Krauss, S. & Hertwig, R. (2000). Muss DNA evidence schwer verständlich sein? Der Ausweg aus einem Kommunikationsproblem [Does DNA evidence need to be difficult to understand?]. *Monatszeitschrift für Kriminologie und Strafrechtsreform, 83*:155–162.

Lander, E. S. (1989). DNA fingerprinting on trial. *Nature, 339*:501–505.

Lander, E. S. & Budowle, B. (1994). DNA fingerprinting dispute laid to rest. *Nature, 371*:735–738.

Lewontin, R. C. & Hartl, D. L. (1991). Population genetics in forensic DNA typing. *Science, 254*:1745–1750.

Lifton, R. J. & Mitchell, G. (2000). *Who owns death? Capital punishment, the American conscience, and the end of executions.* New York: Morrow.

Lindsey, S., Hertwig, R. & Gigerenzer, G. (in press). *Communicating statistical evidence. Jurimetrics.*

Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. *Organizational Behavior and Human Decision Processes, 82*:217–236.

NRC (National research council committee on DNA forensic science: An update) (1996). *The evaluation of forensic DNA evidence.* Washington: National Academy Press.

Politser, P. E. (1984). Explanations of statistical concepts: Can they penetrate the haze of Bayes? *Methods of Information in Medicine, 23*:99–108.

Rattner, A. (1988). Convicted but innocent: Wrongful conviction and the criminal justice system. *Law and Human Behavior, 12*:283–293.

Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science, 253*:980–986.

Scheck, B., Neufeld, P. & Dwyer, J. (2000). *Actual innocence: Five days to execution, and other dispatches from the wrongly convicted.* New York: Doubleday.

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

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.

Stigler, S. M. (1983). 'Who discovered Bayes' theorem?' *The American Statistician, 37*:296–325.

Wells, G. L., Malpass, R. S., Lindsay, R. C. L, Fisher, R. P., Turtle, J. W. & Fulero, S. M. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist, 55*:581–598.

Windeler, J. & Köbberling, J. (1986). Empirische Untersuchung zur Einschätzung diagnostischer Verfahren am Beispiel des Haemoccult-Tests. [An empirical study of the judgments about diagnostic procedures using the example of the hemoccult test.] *Klinische Wochenschrift, 64*:1106–1112.

Appendix

We present the text for the probability and natural frequency versions of one of the two cases involving forensic DNA analysis used by Lindsey *et al.* (in press). Each case description included the testimony of an expert who performed a DNA analysis. The expert testimony provided numerical information about the base rate of the DNA profile, and the analysis' hit rate and false-positive rate. The numerical information was presented in either probabilities or natural frequencies. Participants were asked to estimate two probabilities (or proportions): (1) the probability that a person who is found to match in a DNA analysis (with the evidence from the crime scene) actually has the DNA profile in question, and (2) the probability that a person is the source of the trace recovered from the crime scene given a match in the DNA analysis.

Probabilities

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 be shown to match in this type of DNA analysis is 0.001%. In the present case, the DNA profile of the sample from the defendant matches the DNA profile of the trace recovered from the crime scene.

Natural frequencies

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. Of the 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 the present case ...