Journal of Experimental Psychology: General
2001, Vol. 130, No. 3, 380-400

0096-3445/01/$5.00 DOI: 10.1037//0096-3445.130.3.380

Teaching Bayesian Reasoning in Less Than Two Hours

Peter Sedlmeier
Chemnitz University of Technology

Gerd Gigerenzer
Max Planck Institute for Human Development

The authors present and test a new method of teaching Bayesian reasoning, something about which previous teaching studies reported little success. Based on G. Gigerenzer and U. Hoffrage's (1995) ecological framework, the authors wrote a computerized tutorial program to train people to construct frequency representations (representation training) rather than to insert probabilities into Bayes's rule (rule training). Bayesian computations are simpler to perform with natural frequencies than with probabilities, and there are evolutionary reasons for assuming that cognitive algorithms have been developed to deal with natural frequencies. In 2 studies, the authors compared representation training with rule training; the criteria were an immediate learning effect, transfer to new problems, and long-term temporal stability. Rule training was as good in transfer as representation training, but representation training had a higher immediate learning effect and greater temporal stability.

Statistical literacy, like reading and writing, is indispensable for an educated citizenship in a functioning democracy, and the dissemination of statistical information in the 19th and 20th centuries has been linked to the rise of democracies in the Western world (Porter, 1986). Interest in statistical information such as population figures has been common among political leaders for centuries (e.g., Bourguet, 1987). The willingness to make economic and demographic numbers public rather than to treat them as state secrets, however, is of recent origin: The avalanche of printed statistics after about 1820 both informed the public and justified governmental action to the public (Krüger, Daston, & Heidelberger, 1987). Nevertheless, unlike reading and writing, statistical literacy—the art of drawing reasonable inferences from such numbers—is rarely taught (e.g., Garfield & Ahlgren, 1988; Shaughnessy, 1992). The result of this has been termed "innumeracy" (Paulos, 1988).

In this article, we address the question of how best to teach statistical literacy. We focus on the special case of Bayesian inference with binary hypotheses and binary information (for results of training in reasoning about other kinds of statistical tasks, see Sedlmeier, 1999, 2000). Here are two examples to which this form of statistical inference applies. First, consider the case of a 20-year-old man from Dallas who had a routine HIV test (Gigerenzer, 1998). The test result was positive; the young man assumed this meant he was infected with the virus and was plagued by thoughts of suicide. But what is the probability that he really has the virus given a positive test? Or consider the case of Alan M. Dershowitz, a Harvard professor and advisor to the O. J. Simpson defense team. He stated on U.S. television that only about 0.1% of wife batterers actually murder their wives and claimed that therefore evidence of abuse and battering should not be admissible in a murder trial. But what is the probability that the husband was the murderer, given that he battered his wife and the wife was killed (Good, 1995; Koehler, 1997)?

Bayesian Inference

Our goal is to design an effective method of teaching Bayesian inference. This goal might appear to be doomed to failure for two reasons. First, a large body of experimental results suggests that Bayesian inference is alien to human inference; second, a small number of studies actually attempting to teach people Bayesian reasoning met with little or no success. These two reasons need to be addressed in more detail.

Since the pioneering work of Ward Edwards and his colleagues, an avalanche of experimental studies has investigated whether people reason according to Bayes's rule (for a summary, see Koehler, 1996). Edwards's (1968) major finding was "conservatism," that is, that participants overweighed base rates. In the 1970s, however, Kahneman and Tversky (1972) argued that "in his evaluation of evidence, man is apparently not a conservative Bayesian: he is not a Bayesian at all" (p. 450). Neglect rather than overweighing of base rates became the message of their heuristics-and-biases program in the 1970s and 1980s. "The genuineness, the robustness, and the generality of the base-rate fallacy are matters of established fact" (Bar-Hillel, 1980, p. 215). These demonstra-

Peter Sedlmeier, Department of Psychology, Chemnitz University of Technology, Chemnitz, Germany; Gerd Gigerenzer, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany.

This research was supported by a Feodor Lynen Stipend of the Humboldt Foundation as well as a Habilitationsstipendium of the Deutsche Forschungsgemeinschaft (awarded to Peter Sedlmeier) and a UCSMP Mathematics Project grant from the University of Chicago (awarded to Gerd Gigerenzer).

We are especially grateful to Jim Magnusson and Tom McDougal, who helped the project get started, and to Brad Pasanek, Nicola Korherr, Ursel Dohme, and Gregor Caregnato for their assistance in the training studies. We thank Donna Alexander, Berna Eden, Dan Goldstein, Ralph Hertwig, and Anita Todd for comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Peter Sedlmeier, Department of Psychology, Chemnitz University of Technology, 09107 Chemnitz, Germany, or to Gerd Gigerenzer, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. Electronic mail may be sent to peter.sedlmeier@phil.tu-chemnitz.de or to gigerenzer@mpib-berlin.mpg.de.

tions that human inference deviated radically from Bayesian inference were not confined to laboratory studies; some experts conducted studies in the field and reported similar results. For instance, Eddy (1982) asked physicians to estimate the probability that a woman with a positive mammogram actually has breast cancer, given a base rate of 1% for breast cancer, a hit rate of about 80%, and a false-alarm rate of about 10%. He reported that 95 of 100 physicians estimated the probability that she actually has breast cancer to be between 70% and 80%, whereas Bayes's rule gives a value of about 7.5%. Such systematic deviations from Bayesian reasoning have been called "cognitive illusions," analogous to stable and incorrigible visual illusions (von Winterfeldt & Edwards, 1986; for a discussion of the analogy, see Gigerenzer, 1991).

If the analogy between cognitive illusions and visual illusions holds, the teaching of statistical reasoning should have little hope of success. This conclusion seems to be confirmed by the results of the few studies that have attempted to teach Bayesian inference, using mostly corrective feedback. Peterson, DuCharme, and Edwards (1968) repeatedly showed their participants binomial sampling distributions to correct their "conservative" judgments. Yet, this training did very little to reduce conservatism in further judgments of the same type. Schaefer's (1976) statistically well trained participants received corrective feedback on their estimations of probabilities and also showed practically no training effect. Lindeman, van den Brink, and Hoogstraten (1988) gave corrective feedback on participants' solutions of problems like those used by Kahneman and Tversky (1973). No transfer effect was found in the test phase. Finally, Fong, Lurigio, and Stalans (1990) trained participants on the "law of large numbers" rather than on what they called the "base-rate principle" and thereby only indirectly trained Bayesian inference; this training enhanced the use of base-rate information in only one of several experimental conditions. In these studies, training had little or no success.[1] The negative conclusions of the heuristics-and-biases program (Kahneman & Tversky, 1996) and the meager results of the teaching studies seem to suggest to many what Gould (1992) so bluntly stated: "Tversky and Kahneman argue, correctly I think, that our minds are not built (for whatever reason) to work by the rules of probability" (p. 469).

Bayesian Algorithms Depend on Information Format

In the face of these results there seems to be little hope for a successful method of teaching Bayesian inference and statistical reasoning in general. And we would not have tried had there not been two novel results, both theoretical and empirical (Gigerenzer & Hoffrage, 1995). To understand the novelty of the theoretical results, one needs to recall that research on statistical reasoning has focused on whether cognitive algorithms correspond to the laws of statistics or probability (as Piaget & Inhelder, 1951/1975, claimed for children aged 11 and older) or to simple nonstatistical rules of thumb, as Kahneman and Tversky (1996) claim. However, to discuss human inference only in terms of "what kind of rule?" is incomplete because cognitive algorithms work on information, and information always needs representation (Marr, 1982). Take numerical information and the algorithms in a pocket calculator as an example. Numerical information can be represented by the Arabic system, the Roman system, and the binary system, among others. These representations are mathematically equivalent (an isomorphic mapping exists), but they are not equivalent for a calculator or a mind. The algorithms of pocket calculators are tuned to Arabic numbers as input data and would perform badly if one entered binary numbers. The human mind seems to have evolved and learned analogous preferences for particular formats. Contemplate, for a moment, long division with Roman numerals.

The argument that cognitive algorithms are tuned to particular formats of numerical information connects cognition with the environment and can be applied to Bayesian inference. Assume that some capacity or algorithm for inductive inference has been built up in animals and humans through evolution. To what information format would such an algorithm be tuned? It certainly would not be tuned to percentages and probabilities (as in the typical experiments on cognitive illusions) because these took millennia of literacy and numeracy to evolve as tools of communication. Mathematical probability and percentages are, after all, comparatively recent developments (Gigerenzer et al., 1989). Rather, in an illiterate world, the input format would be *natural frequencies*, acquired by natural sampling (see below).

The crucial theoretical results are (a) that Bayesian computations are simpler when information is represented in natural frequencies compared with probabilities, percentages, and relative frequencies and (b) that natural frequencies seem to correspond to the format of information humans have encountered throughout most of their evolutionary development (Cosmides & Tooby, 1996; Gigerenzer, 1994, 1998; Gigerenzer & Hoffrage, 1995; Kleiter, 1994). Let us illustrate the concept of natural frequencies and how they facilitate computations with the mammography problem introduced earlier, in the form in which it was used in our training study:

A reporter for a women's monthly magazine would like to write an article about breast cancer. As a part of her research, she focuses on mammography as an indicator of breast cancer. She wonders what it really means if a woman tests positive for breast cancer during her routine mammography examination. She has the following data:

> The probability that a woman who undergoes a mammography will have breast cancer is 1%.
>
> If a woman undergoing a mammography has breast cancer, the probability that she will test positive is 80%.
>
> If a woman undergoing a mammography does not have cancer, the probability that she will test positive is 10%.
>
> What is the probability that a woman who has undergone a mammography actually has breast cancer if she tests positive?

[1] There were also attempts to improve Bayesian reasoning by focusing participants' attention on certain parts of Bayes's formula. Fischhoff, Slovic, & Lichtenstein (1979, Study 1) tried to increase participants' sensitivity to the impact of base rates by varying the base rates of a Bayesian problem within the same individual but without giving feedback on participants' solutions. This manipulation had almost no generalizing effect on a second task. In three experiments, Fischhoff and Bar-Hillel (1984) examined the effect of different focusing techniques on performance of Bayesian inference tasks. They found that participants took the information to which the experimenters called their attention into account, but this was done equally for relevant and irrelevant information. In a recent study, Wolfe (1995, Experiment 3) found comparable results.

The numerical information in the mammography problem is represented in terms of single-event probabilities, that is, in a *probability format.* The three pieces of information are the base rate $p(\text{cancer}) = .01$, the hit rate $p(\text{positive} \mid \text{cancer}) = .8$, and the false-alarm rate $p(\text{positive} \mid \text{no cancer}) = .1$. The task is to estimate the *posterior probability* $p(\text{cancer} \mid \text{positive})$. The Bayesian algorithm for computing the posterior probability from the probability format amounts to solving the following equation:

$$p(\text{cancer}|\text{positive}) = \frac{p(\text{cancer})p(\text{positive} \mid \text{cancer})}{p(\text{cancer})p(\text{positive} \mid \text{cancer}) + p(\text{no cancer})p(\text{positive} \mid \text{no cancer})}$$

$$= .01 \times .80/(.01 \times .80 + .99 \times .1)$$

$$= .075. \quad (1)$$

Both laymen and physicians have great difficulties with Bayesian inference when information is given in a probability format (e.g., Abernathy & Hamm, 1995; Dowie & Elstein, 1988). For instance, Hoffrage and Gigerenzer (1998; Gigerenzer, 1996) tested 48 physicians on four standard diagnostic problems, including mammography. When information was presented in terms of probabilities, only 10% of the physicians reasoned consistently with Bayes's rule. Gigerenzer, Hoffrage, and Ebert (1998) studied how AIDS counselors explain what a low-risk client's chances are that he actually has the virus if he tests positive. As an assumed client, one of the authors visited 20 public health centers in Germany to have 20 counseling sessions and HIV tests. All the counselors communicated the risks in probabilities and percentages (rather than in natural frequencies, see below) and consistently overestimated the posterior probabilities of having the virus given a positive test (15 of 20 counselors estimated the probability as 99.9% or higher, whereas a reasonable estimate is about 50%), and some counselors even gave inconsistent probability judgments without noticing.

Do these and similar results imply that people are not Bayesians? As the pocket calculator example illustrates, such a conclusion may be unwarranted. Let us now change the format of information from probabilities and percentages to natural frequencies. Natural frequencies represent numerical information in terms of frequencies as they can actually be experienced in a series of events. More technically, natural frequencies are frequencies that have not been normalized with respect to the base rates; that is, they still carry information about base rates (Gigerenzer & Hoffrage, 1995, 1999):[2]

A reporter for a women's monthly magazine would like to write an article about breast cancer. As a part of her research, she focuses on mammography as an indicator of breast cancer. She wonders what it really means if a woman tests positive for breast cancer during her routine mammography examination. She has the following data:

> Ten of every 1,000 women who undergo a mammography have breast cancer.
>
> Eight of every 10 women with breast cancer who undergo a mammography will test positive.
>
> Ninety-nine of every 990 women without breast cancer who undergo a mammography will test positive.

Imagine a new representative sample of women who have had a positive mammogram. How many of these women would you expect to actually have breast cancer?

What is the Bayesian algorithm when the information is presented in natural frequencies? There are 8 women with positive tests and breast cancer (P & C) and 99 women with positive tests and no breast cancer. Thus, the proportion of women with breast cancer among those who test positive is 8 out of 107 (8 + 99). Expressed in probabilities one gets

$$p(\text{cancer}|\text{positive}) = \#(\text{P \& C})/\#\text{P}$$

$$= 8/107$$

$$= .075. \quad (2)$$

Thus, Bayesian computations are simpler when the information is represented in a frequency format (i.e., natural frequencies) rather than in a probability format (Gigerenzer & Hoffrage, 1995). In the frequency format, one can immediately "see" the answer: About 8 of 107 women who test positive will have cancer. The general point here is that Bayesian algorithms are dependent on the information format. Note that the two information formats—probability and frequency—are mathematically equivalent, and so are the two equations; but the Bayesian algorithms are not computationally and psychologically equivalent.

Consistent with the theoretical result that Bayesian algorithms are simpler to use with natural frequencies than with the widely used probabilities, and the ecological thesis that, if the mind has evolved Bayesian algorithms, these are likely to be tuned to natural frequencies, experimental studies have shown that people are more likely to use Bayesian reasoning with natural frequencies. Gigerenzer and Hoffrage (1995) tested laypeople on 15 Bayesian inference problems such as the mammography problem and found that, in every single one, Bayesian reasoning occurred more often when probabilities were replaced with natural frequencies (the two formats shown earlier), with an average increase in Bayesian solutions from 16% to 46%. Bayesian reasoning was measured both by process analysis and by outcome analysis. Similar results with laypeople were found by Christensen-Szalanski and Beach (1982) and Cosmides and Tooby (1996). Hoffrage and Gigerenzer (1998) tested physicians with an average of 14 years professional experience and found that natural frequencies improve "insight" in physicians to about the same extent as in laypeople. As mentioned earlier, with probabilities, physicians found the Bayesian answer in only 10% of the cases; when the same information was represented in natural frequencies, this number went up to 46%.

We applied these theoretical and empirical results when designing a tutorial program for teaching Bayesian reasoning, focusing on everyday situations rather than on the abstract world of "urns and balls."

---

[2] Natural frequencies must not be confused with frequencies that have been normalized with respect to the base rates. For instance, the information in the mammography problem can be expressed in relative frequencies that are normalized with respect to the base rates: a base rate of .01, a hit rate of 0.80, and a false positive rate of 0.10. Also, absolute frequencies can be normalized: a base rate of 1 in 100, a hit rate of 80 in 100, and a false positive rate of 10 in 100. Normalized frequencies, like probabilities or percentages, are normalized numbers that no longer carry information about natural base rates (e.g., about the base rate of breast cancer). They do not facilitate Bayesian reasoning.

Teaching Bayesian Inference

Teaching representations is an alternative to the traditional program of teaching rules, that is, teaching rules without simultaneously teaching representations (e.g., Arkes, 1981). A rule training program would try to teach Bayesian reasoning by first explaining Bayes's rule in its abstract form and then explaining how to insert single-event probabilities into the rule (Falk & Konold, 1992). We are not aware of any studies on rule training for Bayesian reasoning, but rule training programs exist for other statistical rules, such as for the "law of large numbers," more precisely, for recognizing the impact of sample size (see Sedlmeier & Gigerenzer, 1997, 2000).[3] For instance, Fong and Nisbett (1991) proposed rule training for the law of large numbers and found moderate improvement over an untrained control; when generalization to a new domain was tested after 2 weeks, this moderate effect was considerably diminished (Ploger & Wilson, 1991; Reeves & Weisberg, 1993).

We propose an alternative method: teaching Bayesian reasoning by showing people how to construct frequency representations. For this purpose, we designed two versions of frequency representations. One, the *frequency grid*, has been suggested as a means to make the understanding of statistical tasks easier (e.g., Cole, 1988), and the second, the *frequency tree*, is a variant of a tree structure often used in decision analysis. In the frequency grid tutorial, participants learned how to construct frequency representations by means of grids, and in the frequency tree tutorial, they learned to construct frequency representations by means of trees. We also designed a rule training tutorial as a control, with which participants were taught how to insert probabilities into Bayes's formula. All three tutorials were implemented as a computer program on Macintosh computers, written in Macintosh Common Lisp (Apple Computer, Inc., 1992).

In all conditions, the basic training mechanism was to have participants translate the information in the problem text into a given format, that is, Bayes's formula, the frequency grid, or the frequency tree, and have them practice with those formats. The training procedure for each of the tutorials had two parts. The first part guided participants through two inferential tasks—the sepsis problem (see below) and the mammography problem. In the rule training tutorial, participants were instructed how to insert probability information into Bayes's formula. In the two tutorials that taught frequency representations, the system showed participants how to translate probability information into either a frequency grid or a frequency tree. After they were guided through each step in Part 1, the second part of the training required participants to solve eight additional problems on their own with step-by-step feedback. The system asked them to solve each step before going on to the next one. If participants had difficulties with following the requests or made mistakes, the system provided immediate help or feedback. If, for instance, the user was required to enter numbers in the formula or the frequency tree, and the numbers entered were not correct, the system gave immediate feedback. The user always had a choice between trying again or letting the system perform the corrections. If the user decided to try again, the system supplied some hints that were specific to the format used. If, after several corrective interventions, the user was still unable to fill in the numbers correctly and did not want to try again, the system inserted the correct numbers into the respective nodes (frequency tree) or slots (formula). For all training procedures, the help was sufficient to ensure that all participants would solve all problems correctly and complete the training.

We now describe the rule training procedure and the two frequency representation training procedures (see Sedlmeier, 1997, for a detailed description of an extended version of the system, and for a program that provides a comprehensive treatment of basic probability theory and that includes Bayesian reasoning as a part, see Sedlmeier & Köhlers, 2001).

Rule Training

During training and in all three tutorials, participants saw three windows on the screen. The *problem window*, located in the top right portion (see Figure 1), displayed the problem text, in this case, the text of the sepsis problem. The *tutor window* (white area) provided the explanations and instructions and asked the user to perform certain actions. The *representation window* (left half of Figure 1) performed demonstrations and allowed the user to manipulate its contents. Figure 1 shows a screen at the beginning of the first part of the rule training procedure. Just before, the program had explained to the participant that Bayes's formula allows one to calculate the probability that the walk-in patient who displays the symptoms mentioned in the problem has sepsis. The program had also mentioned that to calculate that probability, one needs p(H), p(not H), p(D | H), and p(D | not H). H is short for hypothesis, such as sepsis, and D stands for data such as the presence of the symptoms. At the current point in the training, the system begins to explain how to extract the numerical information from the problem text. The tutor window in Figure 1 explains which information in the problem text corresponds to p(H), the base rate. In the next step (not shown), the base-rate information is "translated" into a component of Bayes's formula. In this step, the empty slot in the representation window is filled with the base-rate value of 0.1, and it is explained how the next piece of information, p(not H), is calculated from the value of p(H) by subtracting it from 1. Then, analogously, the program explains which parts of the problem text correspond to p(D | H), the hit rate, and p(D | not H), the false-alarm rate, and inserts the respective probabilities, that is, 0.8 (the probability of the symptoms given sepsis) and 0.1 (the probability of the symptoms given no sepsis). When the slots for the four bits of information are filled, the system creates an initially empty "frame" for Bayes's formula and demonstrates how the probabilities are to be inserted into the frame. Inserting the correct numbers into that frame and calculating the result gives the posterior probability p(sepsis | symptoms). Figure 2 shows this final state of the translation process for the mammography problem. All pieces of information needed in Bayes's formula have been extracted from the problem text and inserted into the respective slots (Figure 2, upper left). Then, the frame for Bayes's formula has been filled with the respective probabilities (lower left) and the result has been calculated (lower right). In the second part of the training procedure, the upper part of the representation window, including the formula, was shown immediately. The

[3] Other authors have provided advice on how to reason the Bayesian way but have not reported training studies. Such advice includes structuring the problem, modeling prior probabilities explicitly, stressing the statistical nature of base rate information, clarifying causal chains, and providing individuating information about base rates (e.g., von Winterfeldt & Edwards, 1986).

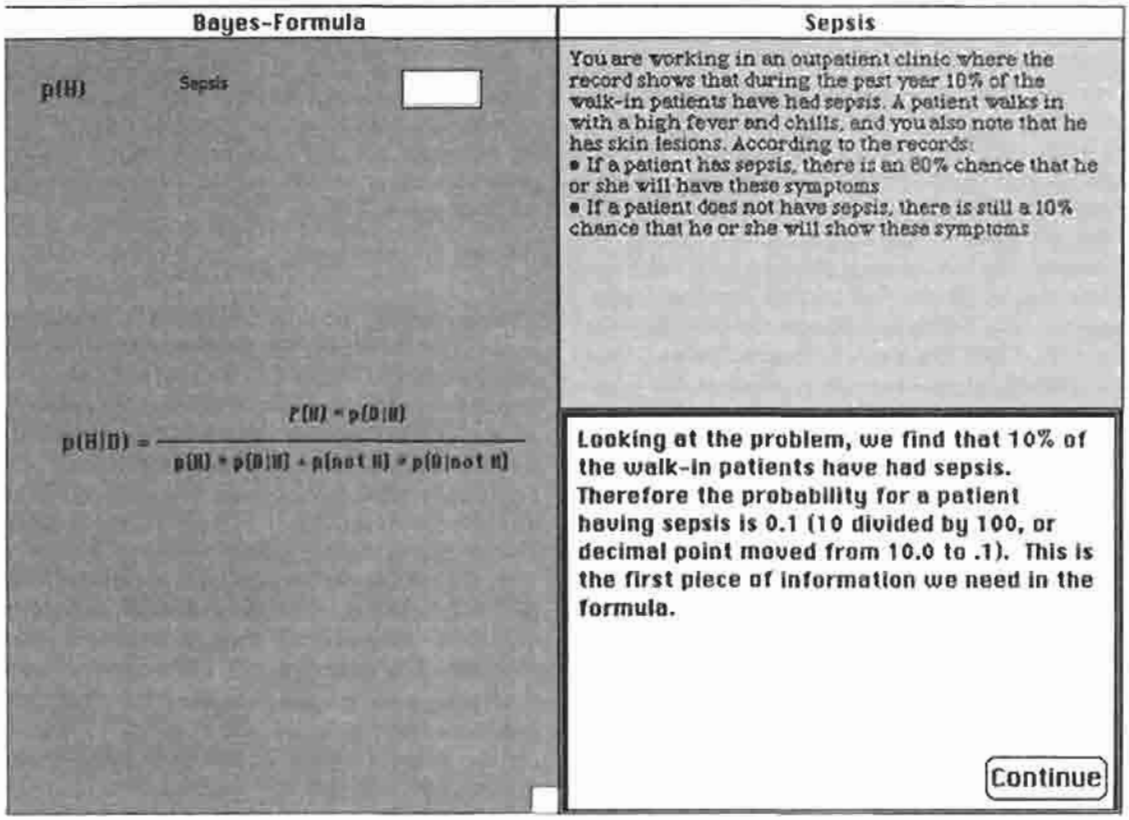

Figure 1. Rule training (Bayes's rule). Screen shot is from the beginning of the first phase of training (sepsis problem).

frame for Bayes's formula appeared only when all the probabilities had been correctly filled in.

Frequency Grid

In a frequency grid, each square represents one case. Figure 3 shows a screen at the beginning of the first part of the frequency grid training procedure. Just before, the program had informed the participants that the empty squares in Figure 3 represent 100 walk-in patients. Again, the tutor window explains which part of the problem text corresponds to the base rate. In the next step (not shown), 10 of the 100 squares are shaded to represent the 10% of walk-in patients who suffer from sepsis. Eventually, circled pluses ("positives") are added to 8 of the 10 shaded squares (corresponding to the hit rate of 80%) and to 9 of the 90 nonshaded squares (corresponding to the false-alarm rate of 10%). Figure 4 shows the point in training when all the information necessary to solve the sepsis problem is filled in on the frequency grid. The ratio of the number of circled pluses in the shaded squares divided by the number of all circled pluses gives the desired posterior probability, that is, p(sepsis | symptoms).

Participants could choose between two grid sizes (100 and 1,000 cases) and were encouraged to select the one that best represented the information given in a problem. For instance, for the mammography problem, the 50 × 20 grid is superior to the 10 × 10 grid because, in the latter, one would have to deal with "rounded" persons. Figure 5 shows the completely filled in frequency grid for the mammography problem, where the ratio of the number of circled pluses in the shaded squares (8) divided by the number of all circled pluses (107) gives the desired posterior probability p(cancer | positive test) = .075.

Frequency Tree

A frequency tree (Figures 6 and 7) does not represent individual cases but constructs a reference class (total number of observations) that is broken down into four subclasses. The top node shows the size of the reference class (100 in Figure 6, and 1,000 in Figure 7), which can be chosen freely in the program. In Figure 6, the program explains how one obtains the base-rate frequency of the walk-in patients to be inserted in the "sepsis" node (left middle node) from the problem text. In the next step (not shown), 10 is inserted in the sepsis node and the program explains how one obtains the number to be inserted in the "no sepsis" node (by subtracting the 10 patients with sepsis from all 100 patients). Eventually, the 10 patients in the sepsis node are divided into 8 (80% of 10) showing the symptoms and 2 not showing the symptoms (left two lower nodes), and the 90 patients in the no sepsis node are divided into 9 (10% of 90) showing the symptoms and 81 not showing the symptoms. The posterior probability p(sepsis | symptoms) is calculated by dividing the number in the left black node, the number of true positives, by the sum of the numbers in both black nodes, the total number of positives.

Figure 7 shows the complete frequency tree for the mammography problem. The two middle nodes specify the base-rate frequencies, that is, the number of cases for which the hypothesis is

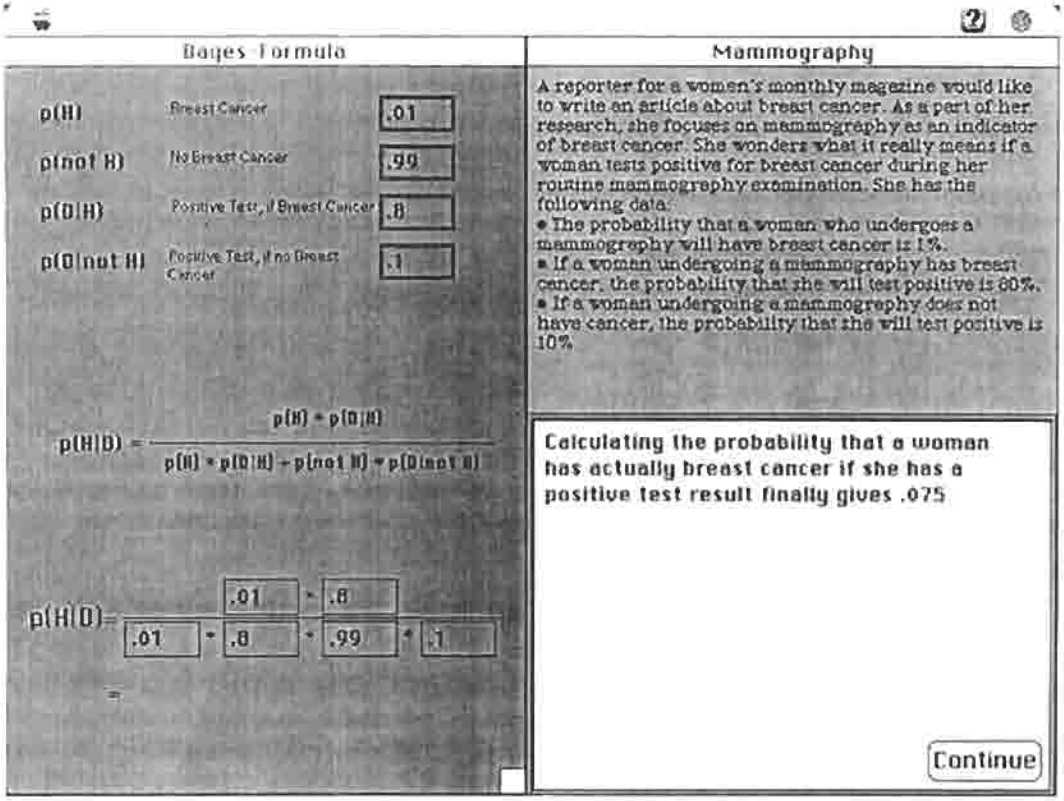

Figure 2. Rule training (Bayes's rule). Screen shot is from the end of the first phase of training (mammography problem).

true (10 women with breast cancer) and the number of cases for which the hypothesis is false (990 women without breast cancer). The four nodes at the lowest level split up the base-rate frequencies according to the diagnostic information (the result of the mammography). The posterior probability *p*(cancer | positive test) is again calculated by dividing the number in the left black node, the true positives, by the sum of the numbers in both black nodes, the total number of positives.

Evaluation of Training Effectiveness

To measure the effect of training—representation or rule training—the test problems were always given in a probability format. Before participants started to work on the problems, the program was explained and it was made sure that they understood all instructions. Figure 8 shows an example of a test problem presented to the participants, the cab problem (Tversky & Kahneman, 1982). The problem text and the question were always in two different windows. Participants did not have to do the calculations; they were encouraged just to type in their solution as a formula. A formula consisted of numbers, arithmetic operators, and parentheses. This answer format was used to minimize errors due to faulty calculations. To avoid a systematic effect of problem difficulty on training results, the order of problems was systematically varied between participants in Study 1a and completely counterbalanced according to a Latin square in Studies 1b and 2.

In all studies, training effectiveness was measured by comparing participants' solution rates immediately after the training (Test 2), about a week after the training (Test 3), and 1–3 months after the training (Test 4) with the solution rates at baseline (Test 1).

Two Scoring Criteria

We used two criteria to classify an answer as a Bayesian solution, one *strict* and one *liberal*. For the strict criterion, the posterior probability calculated by the participant—either in the form of a numerical value or a formula—had to match exactly the value obtained by Bayes's rule, rounding up or down to the next digit (percentage point). This measure might, however, obscure the fact that participants gained some "ballpark" insight that enabled them to produce a sound but inexact response. To take this possibility into account, we also used a more liberal scoring criterion, which counted a participant's estimate as a Bayesian solution when it came within ±five percentage points of the value obtained by Bayes's rule. The liberal criterion, however, increased the possibility that non-Bayesian reasoning is mistaken as Bayesian reasoning. As Gigerenzer and Hoffrage (1995) have demonstrated, participants confronted with Bayesian tasks often use non-Bayesian algorithms, which by accident might yield results that fall into the interval specified by the liberal scoring criterion. The most frequent non-Bayesian algorithms they identified include computing $p(H\&D)$ by multiplying $p(H)$ and $p(D \mid H)$; computing $p(D \mid H) - p(D \mid \text{not } H)$; or simply picking $p(D \mid H)$ or $p(H)$ from the problem description. In the mammography problem, none of these alternative strategies leads to a result that would be misclassified by a liberal scoring rule as a "Bayesian solution," but this

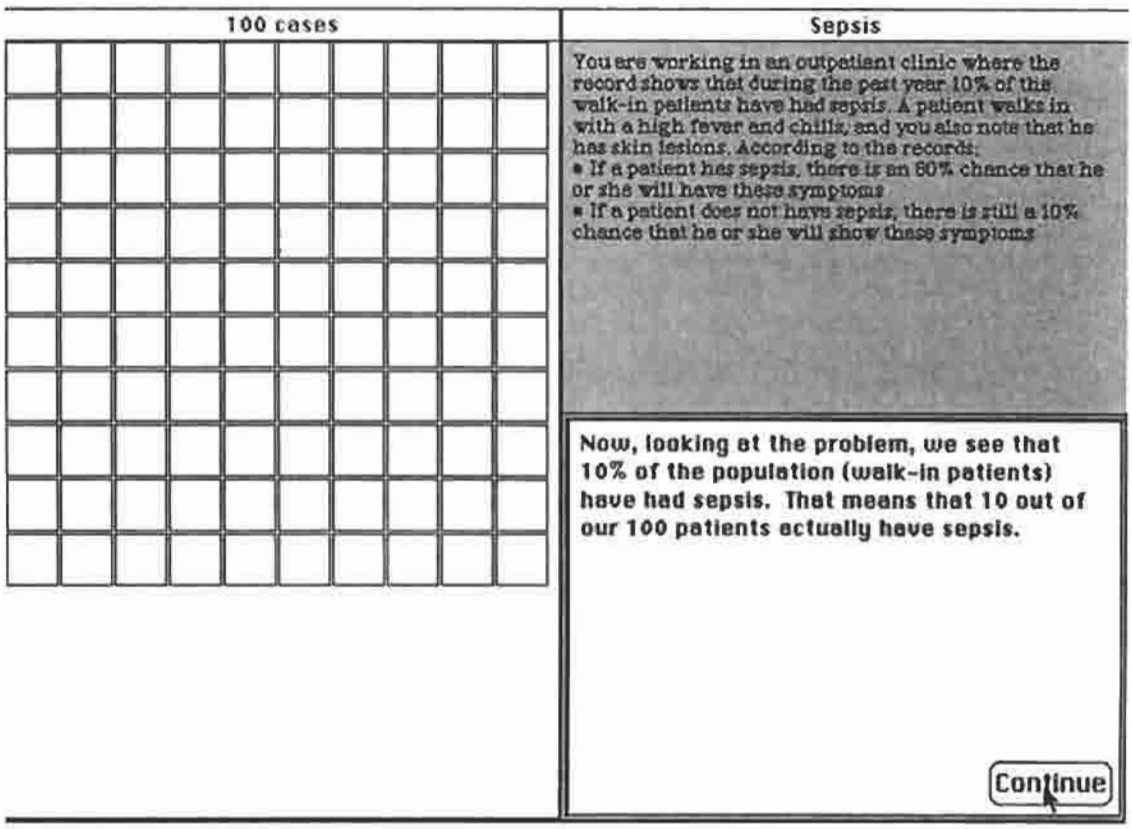

Figure 3. An empty 10 × 10 frequency grid (grid size 100). Screen shot is from the beginning of the first phase of training (sepsis problem).

occurs in other problems. The "rubella problem" illustrates this case:

In Germany, every expectant mother must have an obligatory test for rubella infection because children born to women who have rubella while pregnant are often born with terrible deformities. The following information is at your disposal:

> The probability that a newborn will have deformities traceable to a sickness of its mother during pregnancy is 1%.
>
> If a child is born healthy and normal, the probability that the mother had rubella during her pregnancy is 10%.
>
> If a child is born with deformities and it can be traced to some sickness of the mother, the probability that the mother had rubella during her pregnancy is 50%.
>
> What is the probability that a child will be born with deformities if its mother had rubella during her pregnancy?

The Bayesian solution $p(\mathrm{H} \mid \mathrm{D})$ is .048. But participants who use one of two non-Bayesian algorithms, computing $p(\mathrm{H\&D}) = .005$ or picking $p(\mathrm{H}) = .01$, will produce estimates that lie in the interval of ±5 percentage points around the Bayesian solution. These cases would be misclassified by a liberal scoring criterion but not by a strict scoring criterion (for details, see Gigerenzer & Hoffrage, 1995). To reduce the possibility of such misclassifications, we computed for each problem the results of the non-Bayesian algorithms, and when a participant responded with exactly one of these results, it was counted as an non-Bayesian answer even though it was within the 5 percentage points range.

Three Measures of Training Effectiveness

We measured three possible effects of the training: the immediate training effect (Test 1 compared with Test 2), the generalization or transfer to new problems, and the temporal stability of learning over time (Test 2 compared with Tests 3 and 4). The most interesting measure is stability. Many who teach statistics have the experience that students often study successfully for an exam but quickly forget what they learned after the exam: a steep decay curve. That statistical reasoning does not turn into a habit of mind may not be entirely the students' fault; rather, we conjecture, it is linked to the widespread use of probabilities or percentages as representations for uncertainties and risks. If the thesis is correct that natural frequencies correspond to the format of information humans have encountered throughout most of their evolutionary development, one should expect that decay should not be as quick as with rule training.

Effect sizes rather than significance tests were used for the statistical analysis of training effects (for reasons for using effect sizes, see Cohen, 1990; Loftus, 1993; Rosnow & Rosenthal, 1996; Schmidt, 1996; Sedlmeier, 1996, 1999, Appendix C). Correlational effect sizes (r) in all studies were calculated from the results of significance tests as follows (e.g., Rosenthal & Rosnow, 1991): To evaluate immediate training effects for a given training condition, effect sizes were obtained from repeated measures analyses of variance (ANOVAs) with tests (Test 1, Test 2) as the repeated factors by calculating $r = [F/(F + df)]^{1/2}$. To evaluate differential training effects for two given training conditions, that is, for how

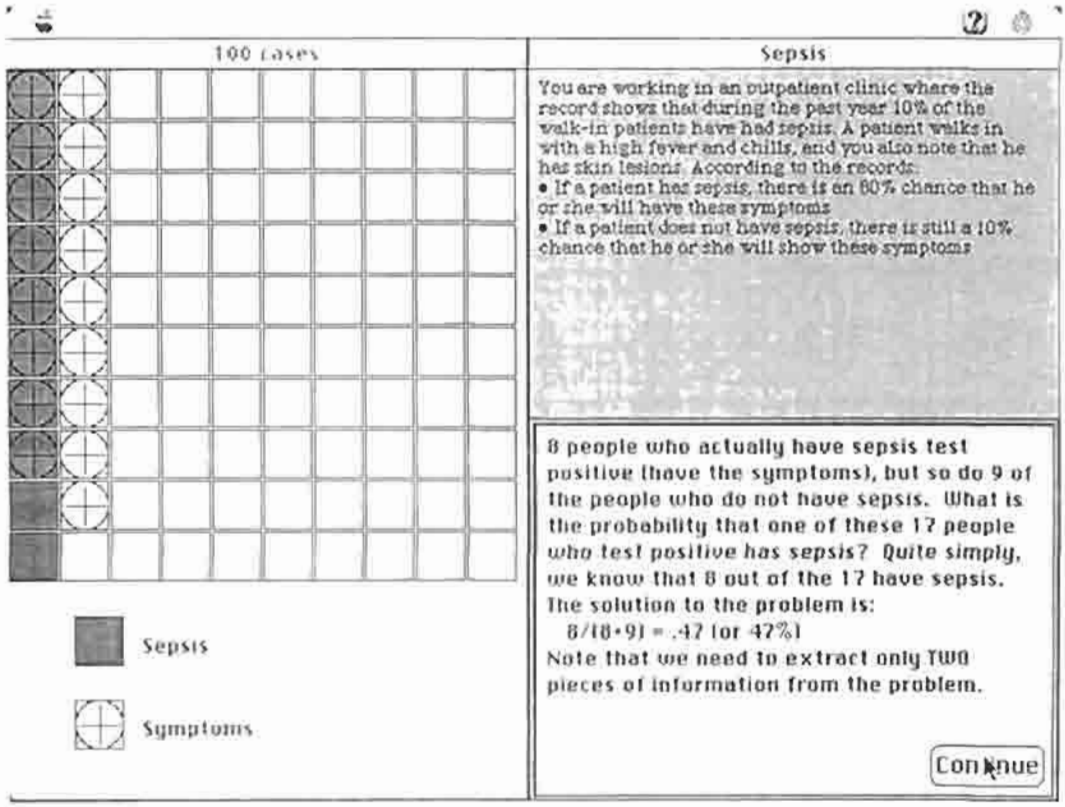

Figure 4. A filled 10 × 10 frequency grid (grid size 100). Screen shot is from the middle of the first phase of training (sepsis problem).

much better one training condition does than the other, the improvement scores for the short-term training effect (Test 2 − Test 1) and for the long-term training effect (Test 4 − Test 1) were used. Effect sizes were obtained from t tests that compared these improvement scores between the two conditions by calculating $r = [t^2/(t^2 + df)]^{1/2}$. The tables that report effect sizes also contain test statistics, that is, values for F and t, and degrees of freedom so that interested readers can easily look up p values from tables of the F and the t distributions.

Note that the effect sizes rely on comparisons between means and therefore rather underestimate the true effects if the distributions contain outliers. This was the case for all studies reported here. Therefore, we report in the figures the more robust medians that can give a more realistic picture. Unless specified otherwise, we report performance in terms of the liberal criterion. Overall, the difference between the two criteria was only one of quantity and not of quality. However, the Appendix shows the complete results, including medians, means, standard deviations, and group sizes for both the liberal and strict scoring criteria.

Study 1a

When people are taught to construct frequency representations, will their Bayesian reasoning improve after training? Will teaching representations enable the transfer of these new skills to new problems? Will performance decay over time or will there be some stability? The computational result (that Bayesian calculations are easier with natural frequencies) and the evolutionary hypothesis (that minds are tuned to frequency representations) gave us some hope for improvement, transfer, and stability. We designed a training study to put our hopes to the test.

Method

Four groups of participants took part in the study. One group worked with the frequency grid, one with the frequency tree, and one with the rule training. A fourth group did not receive training and served as a control. For the three training groups, the study consisted of three sessions with four tests altogether. For the control group, there were two sessions and two tests. The training and all tests were administered on the computer.

Procedure. Test 1 (first session) provided a baseline for performance. Participants were given 10 problems. Before they started to work on the problems, the program was explained and it was confirmed that they understood all instructions. After the baseline test (Test 1), participants in the three training groups received training on 10 problems (2 in Part 1 and 8 in Part 2). They then had to solve another 10 problems (Test 2). The training lasted between 1 and 2 hr; the computerized tutorials allowed participants to work at their own pace. The entire first session (including Tests 1 and 2) lasted between 1 hr 45 min and 3 hr for training groups and between 15 and 30 min for the control group. The second session (1 week after the first session) and the third session (5 weeks after the first session) served to test transfer and stability. Participants in the control group participated in Sessions 1 and 2 only, 1 week apart. In each of the tests, participants had to solve 10 problems, most of them from Gigerenzer and Hoffrage (1995). Two of the problems, the sepsis problem and the mammography problem (see Figures 1 and 2), were used in all four tests and in the training. Tests 3 and 4 each contained one additional "old" problem, that is, a problem already used in the training. All the other problems were

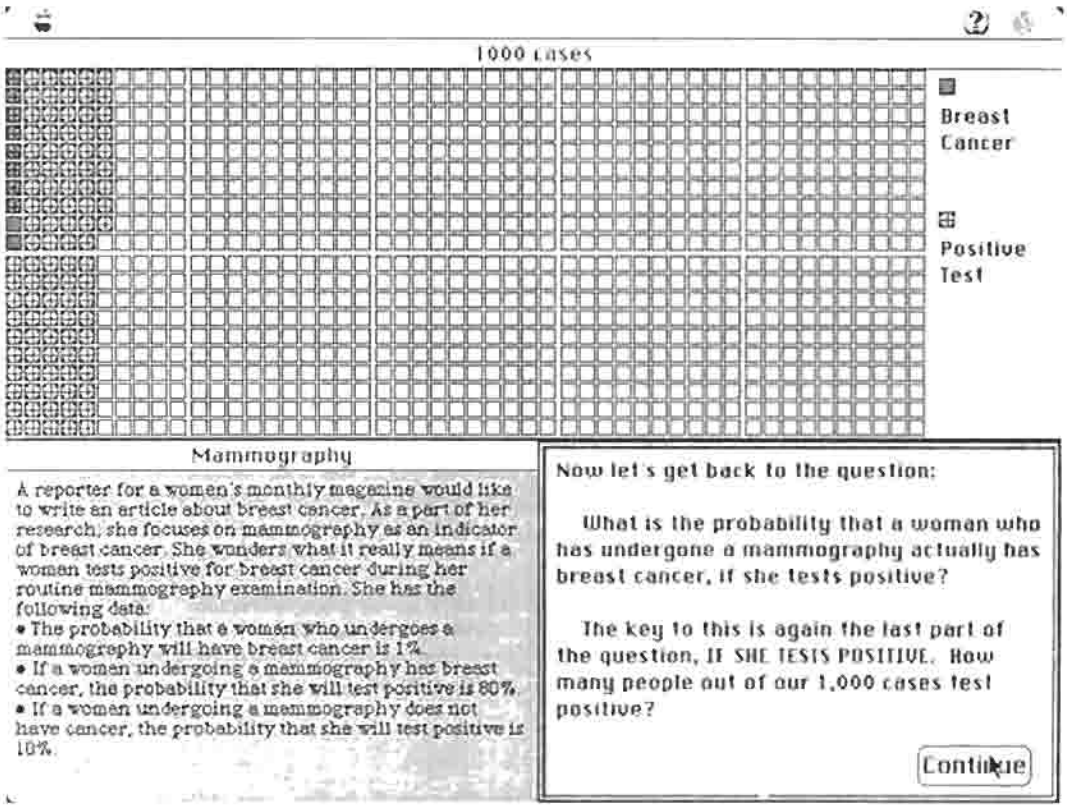

Figure 5. A 50 × 20 frequency grid (grid size 1,000). Screen shot is from the end of the first phase of training (mammography problem).

"new," that is, not used before, either in a test or in the training. The use of both old and new problems allowed us to examine how well the training generalized to problems participants had not seen before. The problems were counterbalanced across sessions, and participants were assigned randomly to one of the four groups.

Participants. Sixty-two University of Chicago students were paid for their participation in two installments, after the first and third sessions, respectively. Six participants who achieved 60% or more correct solutions in Test 1 (baseline) were excluded from the study. Two participants did not complete the first session. We trained 14 participants in the grid condition, 15 participants in the tree condition, 20 participants in the rule training condition, and we had 5 participants in the control condition. We had some loss of participants over the 5-week period due to heavy study loads (end of spring term). The number of participants in the second and third sessions were 12 and 7, respectively, in the grid condition, 13 and 5 in the tree condition, and 15 and 10 in the rule training condition. Four of the five members of the control group took part in the second session.

Results

Figure 9 shows the median percentages of correct solutions for the three training conditions and the control group using the liberal scoring criterion.

Immediate effect. At baseline (Test 1), the median percentage of Bayesian solutions was 10% in the frequency conditions and 0% in the rule training condition. After training, there was a substantial improvement in Bayesian reasoning in each of the three training conditions. The median performance after rule training increased to 60%, whereas it was 75% and 90% for the two frequency representation training sessions. In terms of correlational effect sizes, which express the immediate effect of a training procedure, the training effects were very large for each training, with $r > .90$ for the representation training and $r > .80$ for the rule training (Table 1). In contrast, the control group showed only minimal improvement.

Transfer. To what extent were participants able to generalize from the 10 problems they solved during training (training problems) to problems with different contents (transfer problems)? To test transfer, we compared the solutions for training and transfer problems. Recall that 2 of the training problems, the mammography problem and the sepsis problem, were given in all four tests, and in Tests 3 and 4, participants encountered 1 additional training problem. To the extent that a training method promotes the ability to generalize a technique—to construct frequency representations or to insert probabilities into a formula—there should be little difference between training and transfer problems. A zero value for the difference between training and transfer problems would mean perfect transfer; a large positive value of the size of the difference between training and transfer problems, that is, 60 to 80 percentage points, would mean complete lack of transfer.

The mean percentage of Bayesian solutions was generally almost as high for the transfer problems as for the training problems. With the liberal scoring procedure, the differences between training problems and transfer problems were, on average, 7.2, 3.0, and −0.8 percentage points for the frequency tree, frequency grid, and rule training methods, respectively, and with the strict scoring

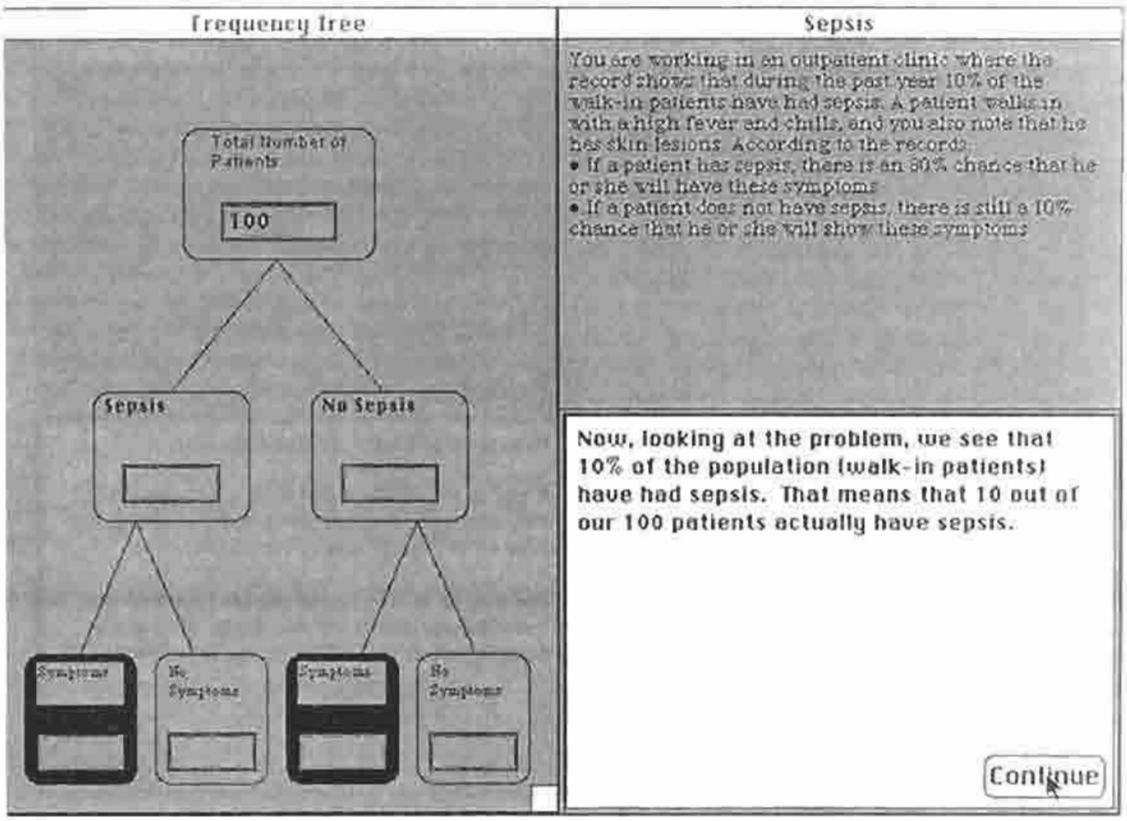

Figure 6. A frequency tree. Screen shot is from the beginning of the first phase of training (sepsis problem).

procedure, they were 6.0, 6.3, and 5.3 percentage points. To summarize, each of the three training programs led to high levels of transfer; that is, participants' average performance in new problems was almost as good as in old problems. Note that this result concerns the difference between the number of Bayesian solutions in training and transfer problems, not the absolute number of Bayesian solutions in transfer problems. The absolute number was consistently larger for those participants who were taught to construct frequency representations (Figure 9).

Stability. For the rule training, Figure 9 shows that, 5 weeks after training, Bayesian reasoning is down to a median of 20%—almost back to where it was before training. The students who were taught to construct frequency representations, however, show a different curve. The higher immediate effect of training is not lost, and 5 weeks after training, there is even an increase in the median number of Bayesian inferences in the frequency grid condition. The calculation of a correlational effect size that expresses the difference in the long-term training effect (Test 4 − Test 1) between the combined frequency conditions and the rule training condition resulted in a medium to large effect size according to Cohen's (1992) conventions (see Table 1, long-term differential training effect).

However, there is possibly an alternative interpretation of the long-term stability: the high attrition rate toward the end of the study. If predominantly weaker participants had dropped out, then the long-term results would be upwardly biased because they would mainly reflect the achievement of the stronger participants. Is there evidence for this conjecture? We checked whether the performance of participants who completed all four tests differed from those who did not. For both representation training versions, the median performance of those who completed all tests was the same as that of the total group shown in Figure 9, except for one point—the frequency grid group at Test 2 matched the median of the frequency tree group. Thus, the results of the representation training seem to be uninfluenced by the attrition. For the rule training, there was a small difference, which is shown in Figure 9. The dotted line shows the performance of those participants who completed all tests. Their performance was slightly above the total group but showed the same pattern of decay. This analysis indicates that the results in Figure 9 are not much influenced by a potential difference between those participants who dropped out and those who completed all four tests.

Discussion

Study 1a showed that all three training programs can improve Bayesian reasoning. The degree of improvement was, as it should be, larger than Gigerenzer and Hoffrage (1995) had reached without training, that is, by merely presenting information in natural frequencies (46% on average, with a strict scoring procedure). The difference between the representation and the rule training was most pronounced in the temporal stability of what participants had learned. The time needed for teaching representations was short, between 1 and 2 hr (not counting the time needed for the tests), depending on the speed of the individual participant.

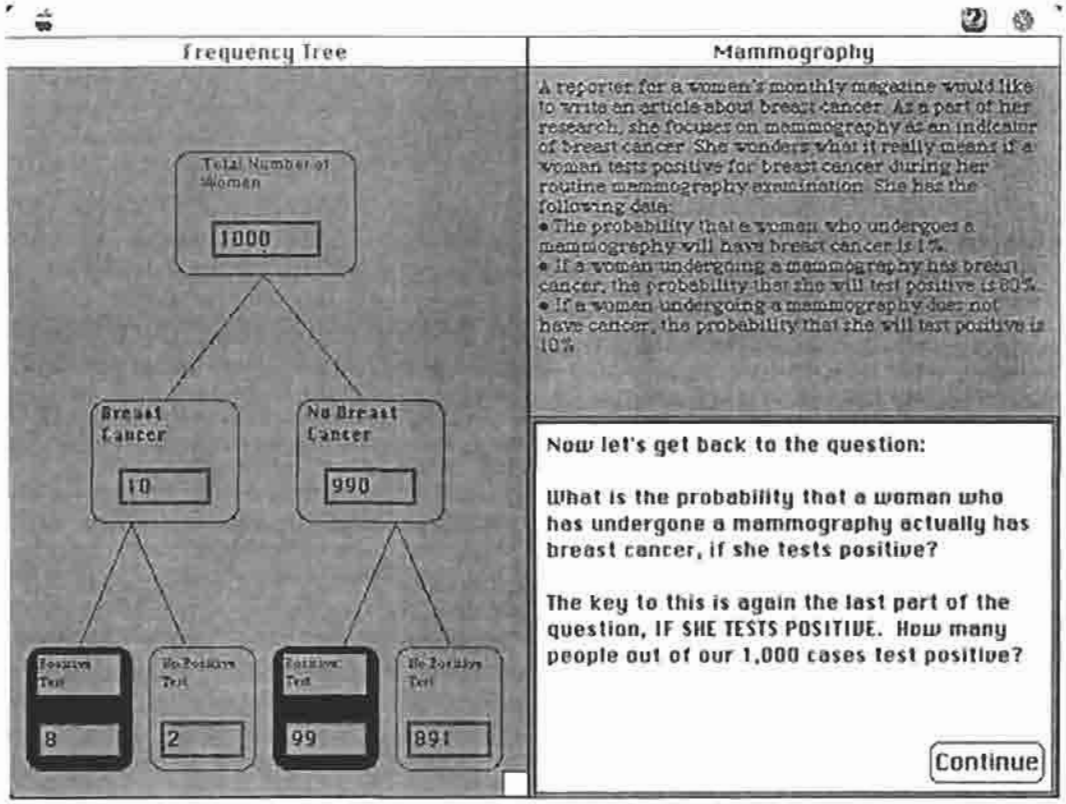

Figure 7. A frequency tree. Screen shot is from the end of the first phase of training (mammography problem).

Study 1a, however, had its limits and therefore should be assigned the status of a pilot study. First, there was the high attrition rate. Although the participants who completed all four tests did not seem to differ from those who completed only the first two or three tests, the high attrition rate may have affected the reliability of the results in Test 4. Second, the study does not necessarily show that the distinction between probabilities and frequencies was the only factor that made a difference because the training conditions differed not only in whether frequency or probability formats were used but also in whether the conditions relied on a graphical aid. Graphics might have been an important factor in achieving training success. Study 1b addressed the first conjecture, and Study 2 addressed the second.

Study 1b

This study investigated whether the results of Study 1a could be replicated in the absence of high attrition rates. To help prevent high attrition rates, participants were paid only at the end of Session 3 rather than in two installments, as in Study 1a. Furthermore, Study 1b addressed the question of whether results are influenced by performance-contingent payment (participants in Study 1a were paid a flat sum, independent of their performance). If a flat fee is paid, participants might not be motivated to do their best (Hertwig & Ortmann, 1999).

Method

Two of the three training programs from Study 1a were used in this study: the frequency tree training and the rule training. German versions of the programs were used because participants in Study 1b were German.

Procedure. Two groups of participants took part in the study. One group was taught with the frequency tree and the other with rule training. About half of the participants in each group were told at the beginning of the first session that the 20% of participants who achieved the best results overall would receive a monetary bonus. The first session contained a baseline test (Test 1), the training, and a posttest (Test 2). Testing and training proceeded as in Study 1a. The German participants took more time than did their American counterparts in Study 1a. Observation of participants suggested that the Germans took the task more seriously than did their American counterparts. If they could not solve a task, they did not easily switch to the next one, a behavior that was frequently observed in the American participants.

To achieve average times comparable to those needed in Study 1a, that is, about 2.5 hr for tests and training combined, the number of tasks was reduced to seven per test, and the number of training tasks was reduced to six (two in Part 1 and four in Part 2). Each test contained two "old" tasks, the sepsis and mammography tasks, that were also used in the training and five "new" tasks, that is, tasks not previously used in either test or training. The second session (Test 3, about 1 week after the first) served to assess transfer and short-term stability. Finally, the third session, which was held, on average, about 5 weeks after the first, measured long-term stability.

Participants. Fifty-six students at the Free University of Berlin, Germany, were paid for their participation. Unlike in Study 1a, none of the participants in Studies 1b and 2 reached more than 60% solutions at the baseline test; thus no participants were excluded in these studies. Twenty-eight participants were trained in each condition. Fourteen participants in the frequency tree condition and 13 in the rule training condition were told that they would receive a monetary bonus if their results were among the best 20%. With the help of the revised payment schedule, there was no attrition of participants over the course of the study.

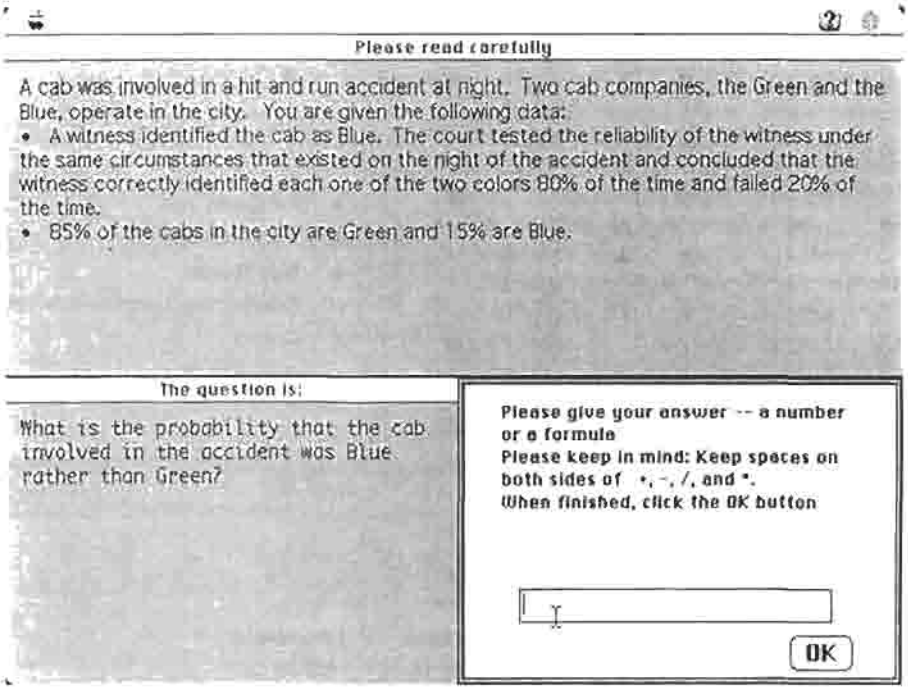

Figure 8. Testing session: The problem text (here, the cab problem) is in the upper window, the question in the lower left window, and the instructions in the lower right window. Participants can type in numbers or formulas consisting of parentheses and basic arithmetic operators.

Results

Figure 10 shows the performance for both training methods (Figure 10a) and this performance broken down to participants with and without the performance-contingent bonus (Figures 10b and 10c).

Immediate effect. Similar to the American participants in Study 1a, the German participants showed little or no skills to solve Bayesian tasks: At Test 1, the median number of problems solved is zero. The immediate training effect (Test 2 – Test 1) is of very similar magnitude as for the American participants in Study 1a: Both teaching methods improve Bayesian reasoning, with a median of 64% for the rule training and 86% for the representation training. This differential effect was more pronounced for participants who were not told about a bonus than for those who could expect to earn one (compare Figures 10b and 10c). The effect size analysis that relied on the more conservative means rather than the medians gives a similar picture (Table 2). All effect sizes measuring the immediate effect were large and were more pronounced in the frequency tree condition than in rule training, in particular when payment was not performance contingent (Table 2, short-term differential training effect). In no single test, regardless of whether there was a prospect of a bonus, did the rule training performance surpass that of the representation training.

Transfer. To test transfer, the solutions in the two "old" tasks that were used in all tests as well as in the training were compared with the results in the "new" tasks that were used only once. Transfer was excellent in both training programs. The average difference between old and new tasks was only 2.4 percentage points in the frequency tree condition and zero in the rule training condition, with the liberal scoring criterion (with the strict scoring criterion, the corresponding values were 4.6 and 2.4 percentage points, respectively). Transfer was not influenced by whether a bonus could be expected. The difference between old and new tasks in the bonus and no-bonus subgroups differed from those in the overall analysis by, at most, 0.9 percentage points.

Stability. Figure 10a shows basically the same pattern shown in Figure 9, except that the decay in the rule training is not as strong—after 5 weeks, performance is down to only 43% compared with 20% (Figure 9). But this direct comparison between the two studies would be misleading because it aggregates over the bonus and no-bonus groups in Study 1b, which show different performance patterns (Figures 10b and 10c). When participants could not win a bonus, the performance was almost identical with that in Study 1a, where participants were also not offered bonuses. The decay curve found in the rule training condition of Study 1a could be replicated almost perfectly (compare Figure 9, "Frequency Tree" and "Rule Training," with Figure 10b): After 5 weeks, a median of only 14% Bayesian solutions was found. If, however, participants had the prospect of a bonus, there was no decay in the rule training condition (see Figure 10c). In contrast, the results in the frequency tree condition were not influenced by whether participants could expect to receive a bonus. For instance, in the no-bonus group, the performance remained at a median of 86% Bayesian solutions over 5 weeks, from Test 2 to Test 4.

Taken together, the no-bonus group provides an almost exact replication of the results found in Study 1a. Also, the effect sizes for the long-term differential training effect, that is, how much better participants learned in the frequency tree than in the rule training condition, are comparable to the one found in Study 1a

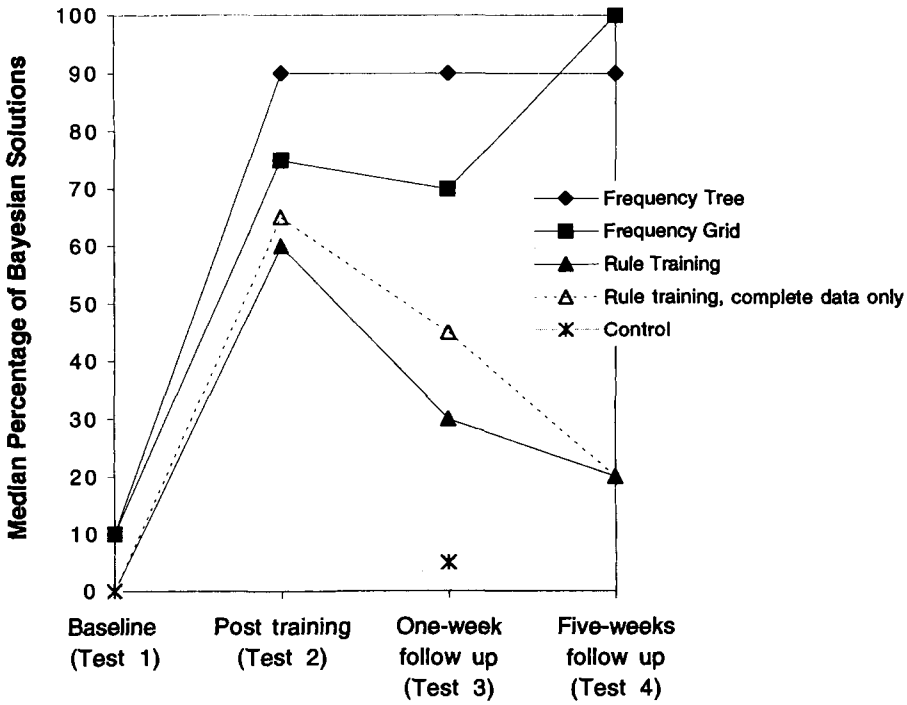

Figure 9. Median percentages of Bayesian solutions obtained in Study 1a (out of 10 possible) for the three training conditions and the control condition (liberal scoring criterion). For the rule training, values for all participants and those participants who completed all four tests are shown separately.

(Table 2, long-term differential training effect, "No bonus"). The combination of a bonus with rule training, however, led to a new result. We try to explain this result in the next section.

Discussion

Study 1b reduced the attrition rate to zero and replicated the major findings of Study 1a: Both the representation and the rule training led to a substantial and immediate improvement in Bayesian inference, with about the same advantage for representation training as in Study 1a; both types of training were equally excellent in transfer; and the representation training provided temporally stable improvements, whereas the rule training showed decay. This holds true for both the median solution rates and the effect sizes based on the means. The new finding was that the rule training did not show a decay when participants were offered a bonus (there was none in Study 1a).

What could be the reason for this bonus effect? We suggest the following: Many German high school students and most German university students have heard about Bayes's formula. At minimum, our participants probably knew where they could find out

Table 1
Correlational Effect Sizes Expressing Immediate Training Effects Within Conditions and Differential Training Effects Across Conditions in Study 1a

	Liberal scoring		Strict scoring		
Training effect	Test statistic	*r*	Test statistic	*r*	*df*
Immediate (Test 2 − Test 1)					
Frequency tree	82.04	.92	84.79	.93	14
Frequency grid	81.56	.93	69.44	.92	13
Rule training	44.39	.84	21.39	.73	19
Short-term differential (Test 2 − Test 1)					
Frequency conditions versus rule training	1.04	.15	2.40	.33	47
Long-term differential (Test 4 − Test 1)					
Frequency condition(s) versus rule training	1.95	.40	1.86	.38	20

Note. The effect sizes for the immediate training effects were calculated from repeated measures ANOVAs with tests (Test 1, Test 2) as the repeated factors, and differential training effects were calculated from *t* tests of group differences using improvement scores (Test 2 − Test 1 and Test 4 − Test 1). The table includes data for both liberal and strict scoring criteria. For each comparison, it shows test statistic (*F* for immediate effects and *t* for differential effects), correlational effect size *r*, and *df*.

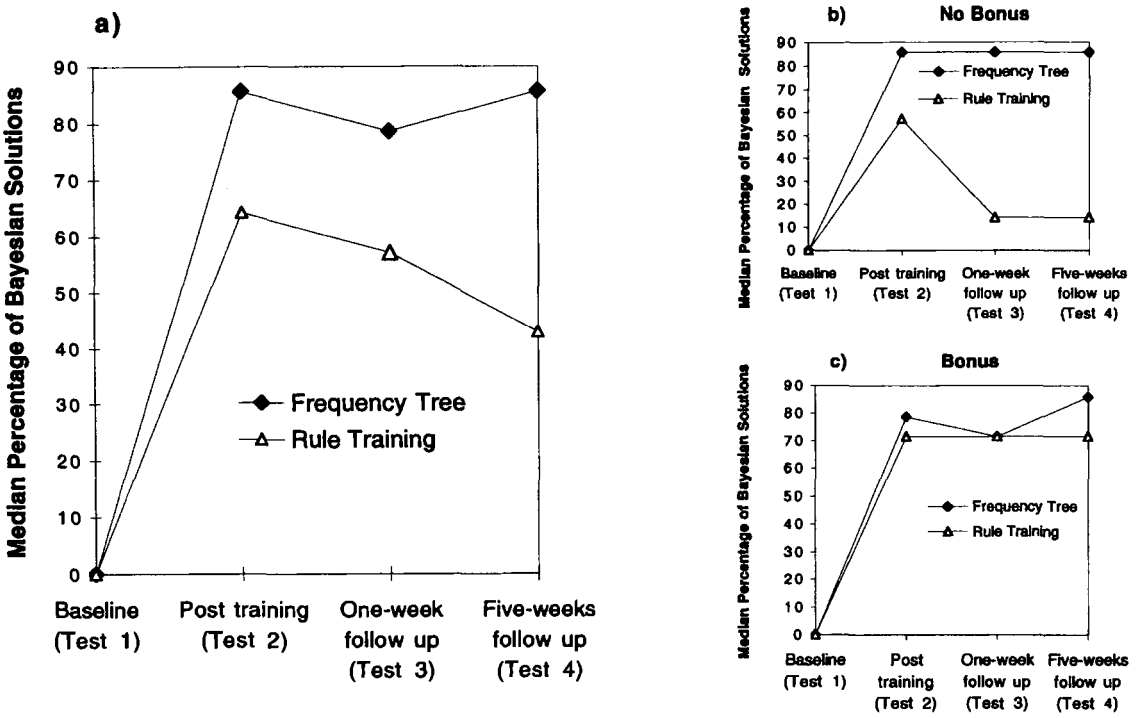

Figure 10. Median percentages of Bayesian solutions obtained in Study 1b (out of seven possible) for the two training conditions (liberal scoring criterion). Combined results (Panel a) and separate results for bonus and no-bonus subgroups (Panels b and c) are shown.

about the formula: in mathematics school books for Grades 10 to 13 and in statistics textbooks. Thus, some of the participants who were motivated by the prospect of a bonus may have looked up Bayes's rule in the books. To check this hypothesis, we tried to contact all 13 participants in the rule training who were told about the bonus. Because of address changes and other reasons, we were able to reach only 7 participants. Of these, only 1, a law student, reported that he had learned the formula during the training and remembered it well over the whole period without thinking much about it. The other 6 conceded that they had recognized the

Table 2
Correlational Effect Sizes Expressing Immediate Training Effects Within Conditions and Differential Training Effects Across Conditions in Study 1b

	Liberal scoring		Strict scoring		
Training effect	Test statistic	*r*	Test statistic	*r*	*df*
Immediate (Test 2 − Test 1)					
Frequency tree	90.34	.88	71.16	.85	27
Bonus	35.61	.86	33.07	.85	13
No bonus	54.56	.90	35.51	.86	13
Rule training	41.04	.78	39.38	.77	27
Bonus	26.24	.83	25.15	.82	12
No bonus	16.00	.72	15.34	.71	15
Short-term differential (Test 2 − Test 1)					
Frequency tree versus rule training	1.65	.22	1.63	.22	54
Bonus	0.70	.14	0.83	.16	25
No bonus	1.57	.29	1.39	.26	27
Long-term differential (Test 4 − Test 1)					
Frequency tree versus rule training	1.98	.26	2.58	.33	54
Bonus	0.41	.08	0.62	.12	25
No bonus	2.47	.43	3.20	.52	27

Note. The effect sizes for the immediate training effects were calculated from repeated measures ANOVAs with tests (Test 1, Test 2) as the repeated factors, and differential training effects were calculated from *t* tests of group differences using improvement scores (Test 2 − Test 1 and Test 4 − Test 1). For each comparison, the test statistic (*F* for immediate effects and *t* for differential effects), correlational effect size *r*, and *df* are shown.

formula from some statistics course and had thought about it before the retests. Two of the participants said that they had looked it up in statistics textbooks, and I admitted to having made a copy of the formula from the training session and practicing with that copy at home. Thus, it seems that additional effort is a plausible explanation for the better results of those participants in the rule training condition who could expect to receive a bonus.

In contrast, there was no way that participants could learn about the frequency tree because German mathematics or statistics textbooks do not introduce that kind of representation. This interpretation suggests that financial incentives can play an important role in statistical training, in leading to additional efforts to look up the formula outside the laboratory. Nobody denies that students can, in principle, learn to apply Bayes's rule successfully (otherwise, there would be no experts in statistics), and this study has shown that monetary incentives help. Frequency representations, however, still lead to slightly better (and cheaper) results without a monetary motivation.

Study 2

Study 1b successfully replicated the difference in learning effect between a training program using a frequency representation and one relying on the use of rules. However, the results of both Studies 1a and 1b still left open an alternative explanation for the superiority of the representation training over the rule training: Perhaps it was not the difference between frequency and probability formats, but rather whether graphical aids were used, that was responsible for the difference in training results. The main aim of Study 2 was to test this objection. Furthermore, this study examined whether the stability over time found for the representation training in the previous studies holds for a longer period of time—15 weeks rather than 5 weeks. To avoid the potential influence of looking up Bayes's rule outside the laboratory, no bonus was offered in this study. As in Study 1b, participants were paid at the end of Session 3.

A graphical aid, the tree, was used for both the frequency and the probability formats. If the graphical aid is the decisive factor in Bayesian inference training, then there should be no systematic difference in results between the probability and frequency conditions. The tree conditions were compared against the standard rule training used in the previous studies.

Method

Two of the three training methods used in this study, the rule training and the frequency tree training, were identical to the ones in Study 1b. We refer to the third training method as the "probability tree" training.

Probability tree. In a probability tree (Figure 11), the top node contains the value 1, that is, the probability that the respective hypothesis is true or not true. In the specific example that uses the mammography task (see earlier example), this is the probability that a woman who has undergone a mammography does or does not have breast cancer. The two middle nodes show the base-rate probabilities of breast cancer (p = .01) and its complement, no breast cancer (p = .99). The four nodes at the lowest level split up the base-rate probabilities according to the diagnostic information—in our case, the result of the mammography. Only the values in the two shaded nodes are needed to calculate the posterior probability, p(cancer | positive test), because p(cancer | positive test) = p(cancer & positive test)/p(positive test), where p(cancer & positive test) is represented by the left black node, and the sum of both black nodes gives p(positive test). Thus, calculation in the probability tree is identical to that in the frequency tree except that the value in the top node is always 1.

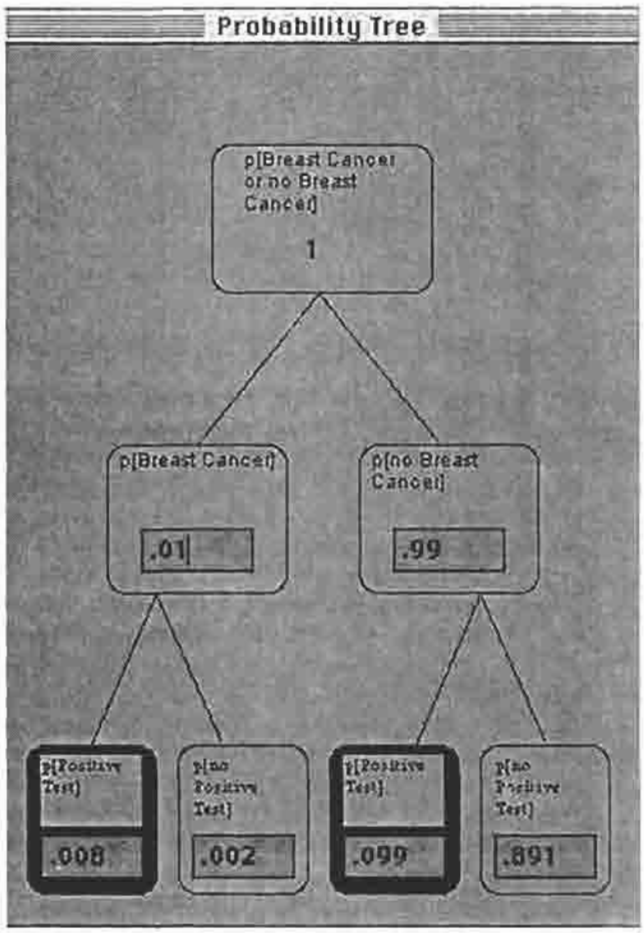

Figure 11. Probability tree as used in Study 2. Screen shot is from the first phase of training (mammography task).

Procedure. Three groups of participants took part in the study. One group worked with the frequency tree, the second with the probability tree, and the third with Bayes's formula. Participants used German versions of the tasks from Study 1a and completed three sessions. The first session contained a baseline test (Test 1), the training, and a posttest (Test 2). Testing and training proceeded as in Study 1b. The second session (Test 3, about 1 week after the first) served to assess transfer and short-term stability. Finally, the third session, which was held about 15 weeks after the first, measured long-term stability. The average intervals between training and Test 4 for the frequency tree, the probability tree, and the rule training condition were 15.4 weeks, 14.8 weeks, and 14.8 weeks, respectively. This prolonged time interval allowed us to test to what degree the excellent stability observed in Studies 1a and 1b, 5 weeks after training, still existed at the later time.

Participants. Seventy-two students at the University of Munich, Germany, were paid for their participation. Twenty-four participants were trained in each of the three conditions. The data for one participant in the rule training condition were lost due to a computer breakdown. There was no attrition in the first two sessions, but there was attrition in the third session (Test 4), probably due to the long time interval (14 weeks) between Sessions 2 and 3. The number of participants in Test 4 was n = 21, n = 21, and n = 18 in the frequency tree, probability tree, and rule training conditions, respectively.

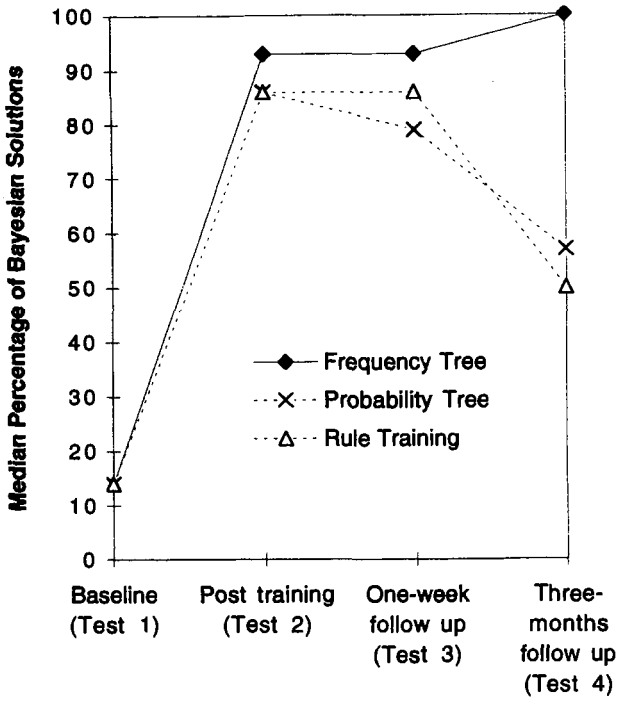

Figure 12. Median percentages of Bayesian solutions obtained in Study 2 (out of seven possible) for the three training conditions (liberal scoring criterion).

Results

The same two scoring criteria as in Studies 1a and 1b were used. Figure 12 shows the median percentages of Bayesian solutions, for the liberal scoring criterion. Again, results for liberal and strict criterion differed in quantity but not in quality.

Immediate effect. As in the previous studies, the baseline test (Test 1) indicated that participants had few skills for solving Bayesian tasks. Before the training, the median percentage of Bayesian solutions over all participants was 14%. The immediate training effect was strong for all three training programs and again yielded large effect sizes that were comparable to those obtained in the previous studies (see Table 3).

Transfer. As in the previous studies, transfer was excellent in all three training programs. On average, using the liberal scoring criterion, the difference between old and new tasks was 0.6 percentage points in the frequency tree, 4.9 in the probability tree, and 4.8 in the rule training condition (using the strict scoring criterion, the corresponding values were 2.1, 3.8, and 1.2 percentage points, respectively).

Stability. In the previous studies, the effect of the representation training was stable over a 5-week period. In rule training, by contrast, the effect faded away over time (with the notable exception of the bonus group in Study 1b). Can participants still maintain their representation skills 15 weeks after training? Do we still obtain the difference between the rule and representation training as in Studies 1a and 1b? Figure 12 shows that, consistent with the results in the previous studies, no decay occurred in the group that received representation training (frequency tree). Here, the immediate training effect of a median of 93% Bayesian solutions remained stable over the whole period of 15 weeks. In fact, it even increased to 100% Bayesian solutions at Test 4. In contrast, the rule training group began high, at a median of 86% Bayesian solutions, and ended up at a median of 50% after 15 weeks. The probability tree training shows a similar pattern of results as the rule training. There is some decay from Test 2 to Test 3 and a more pronounced decay from there to Test 4, with a final level of 57% Bayesian solutions. A comparison of the long-term improvement scores (Test 4 − Test 1) between the frequency tree condition on the one hand and the probability tree and rule training conditions on the other hand again yields medium- to large-sized effects. The difference between the two probability conditions is, by contrast,

Table 3
Correlational Effect Sizes Expressing Immediate Training Effects Within Conditions and Differential Training Effects Across Conditions in Study 2

	Liberal scoring		Strict scoring		
Training effect	Test statistic	*r*	Test statistic	*r*	*df*
Immediate (Test 2 − Test 1)					
Frequency tree	143.98	.93	94.00	.90	23
Rule training	50.98	.84	63.42	.86	22
Probability tree	115.00	.91	187.29	.94	23
Short-term differential (Test 2 − Test 1)					
Frequency tree versus rule training	1.88	.27	0.74	.11	45
Frequency tree versus probability tree	1.05	.15	0	0	46
Probability tree versus rule training	1.00	.15	0.85	.13	45
Long-term differential (Test 4 − Test 1)					
Frequency tree versus rule training	3.11	.46	2.69	.40	37
Frequency tree versus probability tree	2.98	.43	2.02	.30	40
Probability tree versus rule training	0.07	.01	0.77	.13	37

Note. The effect sizes for the immediate training effects were calculated from repeated measures ANOVAs with tests (Test 1, Test 2) as the repeated factors, and differential training effects were calculated from *t* tests of group differences using improvement scores (Test 2 − Test 1 and Test 4 − Test 1). For each comparison, it shows test statistic (*F* for immediate effects and *t* for differential effects), correlational effect size *r*, and *df*.

small, especially when the liberal scoring criterion is used (Figure 12; see Table 3, long-term differential effect).[4]

Discussion

Studies 1a and 1b left open a possible alternative explanation for the superior results in the representation training as compared with the rule training. The former used graphical aids, whereas the latter did not, and therefore the graphical aid might have made the difference. In Study 2, both a frequentistic and a probabilistic condition used the same graphical aid, a tree structure. The immediate training results were very high in both tree conditions, but they differed markedly in the stability of the training success over time. The frequency tree training enabled participants to retain what they had learned more than 3 months before, whereas the effect of the probability tree training decayed over time, to a median of 57%.

How much could the probability training gain by using a graphical aid? Figure 12 shows that performance was slightly better after 15 weeks, but overall, there is little, if any, difference. This holds despite the rule training group having to learn a more complicated formula (Bayes's rule for probabilities) than the probability tree group. The similar performance in the two probability training programs indicates that the important question is not whether a graphical aid should be used in teaching statistical literacy but what is a proper representation for a graphical aid.[5] It also indicates that the superior effect of natural frequencies is not due solely to computational simplicity, which is the same for probability trees as for frequency trees except that the decimal point is moved to the left. The results in Study 2 are consistent with Gigerenzer and Hoffrage's (1995) conclusion that natural frequencies constitute a proper representation of uncertainties.

Conclusion

Gigerenzer and Hoffrage (1995) have stressed the importance of studying cognitive algorithms in tandem with the information format for which they are designed. The thesis is that humans and animals more easily encode information about uncertain environments in terms of natural frequencies compared with probabilities, and one can show that Bayesian computations are simpler when information is represented in natural frequencies. Both the frequency grid and the frequency tree are realizations of natural sampling of frequencies.

We applied Gigerenzer and Hoffrage's work to an unresolved problem: how to design a method for teaching Bayesian reasoning that is built on psychological principles and can overcome the lack of success reported in previous studies. The central idea is to teach people to represent information in a way that is tuned to their cognitive algorithms. Whether such cognitive algorithms are the direct result of evolution or whether they rely on evolved mental architectures and are shaped to a large extent during ontogenesis by learning processes does not matter much for our argument. For instance, Sedlmeier (1999) showed that an associative learning model also arrives at the prediction that Bayesian algorithms crucially depend on information format. Our research emphasizes the role of the information representation at encoding. If information is encoded in terms of natural frequencies, probability judgments can be quite exact (Sedlmeier, 1999, pp. 161–163).

This psychological approach was contrasted with the traditional approach to the teaching of statistical reasoning, which emphasizes how to insert the right numbers into the right rule. Similar to prior training attempts on the impact of sample size (e.g., Fong & Nisbett, 1991), the rule training method showed a substantial short-term increase in performance, and relative to this increase, an excellent transfer. After several weeks, however, Bayesian reasoning had undergone the well-known decay function. When participants were taught representations instead of rules, the initial training effect was noticeably higher, transfer was equally good, and there was no loss of performance after 15 weeks.

Let us reflect on the larger context in which the present approach to teaching stands. First, there is an ecological perspective: Cognitive algorithms (or rules) are adapted to specific information formats in the environment. Specifically, the external representation of information can "perform" part of the computations. Second, there is the evolutionary distinction between the past environment to which the cognitive processes of an organism are adapted and the present environment in which an organism lives (e.g., Buss, Haselton, Shackelford, Bleske, & Wakefield, 1998; Cummins, 1998). When environments change, such as by the invention of new forms for the representation of information such as probabilities, cognitive processes may no longer function as well as before, and "illusions" can be a consequence. As an example from vision, consider color constancy, an impressive adaptation of the human perceptual system. It allows people to see the same color under changing illuminations: under the bluish light of day as well as the reddish light of the setting sun. Color constancy, however, fails under certain artificial lights such as sodium or mercury vapor lamps, which were not present in the environment when mammals first evolved (Shepard, 1992). The same type of argument can be made for statistical reasoning (Gigerenzer, 1998), where natural frequencies correspond to the format of information a foraging organism would have encountered before the invention of books and statistics, and probabilities and percentages correspond to an information environment that has been changed by the invention of mathematical probability.

Compared with the earlier emphasis on demonstrating cognitive biases in statistical reasoning, or so-called "inevitable" illusions (e.g., Piattelli-Palmarini, 1994), the ecological perspective can

[4] There is one notable exception from the finding that the results based on the strict and liberal scoring criteria differ only in quantity, in Study 2. According to the strict scoring criterion, there is a relatively large difference in the median percentages at Test 4 (36 percentage points) between rule training and probability tree conditions, which is much smaller when expressed in means (6 percentage points) and which is not found when applying the liberal criterion (see Figure 12 and the Appendix). The large difference is, however, due in part to the coarse step size that determines the possible median percentage values. Recall that with eight possible values (0 to 7 problems solved) a difference of one problem solved amounts to a difference of 14.3 percentage points.

[5] An alternative way to disentangle the possible influence of a graphical aid from that of the information representation (frequentistic vs. probabilistic) would have been to dispense with graphical aids in both representations (rather than to use them in both representations, as in Study 2). We did not proceed with this route because it has already been shown by Gigerenzer and Hoffrage (1995) that frequency representations yield solution rates about three times as high (about 50% correct solutions) as probability representations—both without graphical aids.

actually advise us how to help people understand statistical information. Here, the external representation of numerical information, and the internal translation of one representation into another, can be a major tool for helping people to attain insight. This is not to say that frequency representations are the only tool. Study 1b, for instance, indicated that offering monetary incentives can motivate students to make additional effort and can enable them to perform about as well as those who had a representation training.

We conclude with some open questions and possible extensions of the present work. First, we have dealt with only an elementary form of Bayesian inference, and we do not know how these results generalize to situations in which hypotheses and data are not binary but multivalued or continuous. Second, we have not dealt with situations in which there is more than one piece of diagnostic information, such as two medical tests in sequence. Multiple pieces of information can be reduced to the sequential application of two frequency representations, and Krauss, Martignon, and Hoffrage (1999) have shown that the effect of natural frequencies remains as strong with two pieces of information as it is with one. This result suggests extending teaching representations to situations with multiple pieces of information. Third, an extension of the training program would be to teach Bayesian shortcuts, as described in Gigerenzer and Hoffrage (1995). For instance, when a disease is rare (low base rate) and can be easily detected (high hit rate) and false positives are numerous, as compared with true positives, then the ratio between base rate and false-alarm rate is a good approximation of the Bayesian estimate. For instance, assume that only 2 out of 10,000 men have HIV; the hit rate of an ELISA test is very high; and there are about 20 false positives among those 9,998 men who do not have the virus. The probability that a man who tests positive actually has the virus can be approximated by simply dividing the base-rate frequency (2) by the false-alarm frequency (20); this shortcut results in a value of 1 in 10. A final extension of the training program would be to teach participants to understand and judge the assumptions for the applicability of Bayes's rule (e.g., Earman, 1992) as well as other, competing statistical methods for inference.

Tutorial programs could play a useful role in education for mathematical and statistical literacy and in overcoming innumeracy (Paulos, 1988; Sedlmeier, 1999, 2000). Because the representation training lasts only 1–2 hr, it can be used, for instance, in high school curricula to teach young people how to evaluate the results of pregnancy, HIV, or drug tests. Similarly, it can be used to teach both patients and physicians to estimate the chances of actually having breast cancer after a positive mammogram, and the like. Computerized programs have been proven to attract the attention of young and old alike, and we have observed in our participants a high degree of involvement and desire to succeed. The teaching of statistical literacy can take advantage of human psychology.

References

Abernathy, C. M., & Hamm, R. M. (1995). *Surgical intuition: What it is and how to get it.* Philadelphia, PA: Hanley & Belfus.

Apple Computer, Inc. (1992). *Macintosh common lisp reference.* Cupertino, CA: Author.

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49,* 323–330.

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

Bourguet, M. -N. (1987). Decrire, compter, calculer: The debate over statistics during the Napoleonic period. In L. Krüger, L. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 305–316). Cambridge, MA: MIT Press.

Buss, D. A., Haselton, M. G., Shackelford, T. K., Bleske, A. L., & Wakefield, J. C. (1998). Adaptations, exaptations, and spandrels. *American Psychologist, 53,* 533–548.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Cole, W. G. (1988). Three graphic representations to aid Bayesian inference. *Methods of Informatics in Medicine, 27,* 125–132.

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

Cummins, D. D. (1998). Social norms and other minds: The evolutionary roots of higher cognition. In D. D. Cummins & C. Allen (Eds.), *The evolution of mind* (pp. 31–50). New York: Oxford University Press.

Dowie, J., & Elstein, A. (1988). *Professional judgment: A reader in clinical decision making.* Cambridge, England: Cambridge University Press.

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory.* Cambridge, MA: MIT Press.

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, England: 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.

Falk, R., & Konold, C. (1992). The psychology of learning probability. In F. S. Gordon & S. P. Gordon (Eds.), *Statistics for the twenty-first century* (pp. 151–164). Washington, DC: The Mathematical Association of America.

Fischhoff, B., & Bar-Hillel, M. (1984). Focusing techniques: A shortcut to improving probability judgments? *Organizational Behavior and Human Performance, 34,* 175–194.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance, 23,* 339–359.

Fong, G. T., Lurigio, A. J., & Stalans, L. J. (1990). Improving probation decisions through statistical training. *Criminal Justice and Behavior, 17,* 370–388.

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General, 120,* 34–45.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal of Research in Mathematics Education, 19,* 44–63.

Gigerenzer, G. (1991). On cognitive illusions and rationality. *Poznan Studies in the Philosophy of the Sciences and the Humanities, 21,* 225–249.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 129–162). New York: Wiley.

Gigerenzer, G. (1996). The psychology of good judgment: Frequency formats and simple algorithms. *Journal of Medical Decision Making, 16,* 273–280.

Gigerenzer, G. (1998). Ecological intelligence: An adaptation for frequen-

cies. In D. E. Cummins & C. Allen (Eds.), *The evolution of mind* (pp. 9–29). New York: Oxford University 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 Mellers and McGraw (1999). *Psychological Review, 106,* 425–430.

Gigerenzer, G., Hoffrage, U., & Ebert, A. (1998). AIDS counselling for low-risk clients. *AIDS CARE, 10,* 197–211.

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, England: Cambridge University Press.

Good, I. J. (1995, June). When batterers turn murderer. *Nature, 375,* 541.

Gould, S. J. (1992). *Bully for brontosaurus: Further reflections in natural history.* New York: Penguin Books.

Hertwig, R., & Ortmann, A. (1999). Experimental practices in economics: A methodological challenge for psychologists? Manuscript submitted for publication.

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

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.

Kleiter, G. (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. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavior and Brain Sciences, 19,* 1–54.

Koehler, J. J. (1997). One in millions, billions, and trillions: Lessons from *People v. Collins* (1968) for *People v. Simpson* (1995). *Journal of Legal Education, 47,* 214–223.

Krauss, S., Martignon, L., & Hoffrage, U. (1999). Simplifying Bayesian inference. In L Magnani, N. Nersessian, & N. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 25–31). New York: Plenum Press.

Krüger, L., Daston, L., & Heidelberger, M. (Eds.). (1987). *The probabilistic revolution: Vol. 1. Ideas in history.* Cambridge, MA: MIT Press.

Lindeman, S. T., van den Brink, W. P., & Hoogstraten, J. (1988). Effect of feedback on base-rate utilization. *Perceptual and Motor Skills, 67,* 343–350.

Loftus, G. R. (1993). A picture is worth a thousand *p* values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments & Computers, 25,* 250–256.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: Freeman.

Paulos, J. A. (1988). *Innumeracy: Mathematical illiteracy and its consequences.* New York: Vintage Books.

Peterson, C. R., DuCharme, W. M., & Edwards, W. (1968). Sampling distributions and probability revision. *Journal of Experimental Psychology, 76,* 236–243.

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

Piattelli-Palmarini, M. (1994). *Inevitable illusions. How mistakes of reason rule our minds.* New York: Wiley.

Ploger, D., & Wilson, M. (1991). Statistical reasoning: What is the role of inferential rule training? Comment on Fong and Nisbett. *Journal of Experimental Psychology: General, 120,* 213–214.

Porter, T. M. (1986). *The rise of statistical thinking 1820–1900.* Princeton, NJ: Princeton University Press.

Reeves, L. M., & Weisberg, R. W. (1993). Abstract versus concrete information as the basis for transfer in problem solving: Comment on Fong and Nisbett (1991). *Journal of Experimental Psychology: General, 122,* 125–128.

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

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1,* 331–340.

Schaefer, R. E. (1976). The evaluation of individual and aggregated subjective probability distributions. *Organizational Behavior and Human Performance, 17,* 199–210.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen [Beyond the ritual of significance testing: Alternative and supplementary methods]. *Methods of Psychological Research-online, 1.* Available on the World Wide Web: http://www.mpr-online.de/.

Sedlmeier, P. (1997). BasicBayes: A tutor system for simple Bayesian inference. *Behavior Research Methods, Instruments, & Computers. 29,* 328–336.

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

Sedlmeier, P. (2000). How to improve statistical thinking: Choose the task representation wisely and learn by doing. *Instructional Science, 28,* 227–262.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers? *Journal of Behavioral Decision Making, 10,* 33–51.

Sedlmeier, P., & Gigerenzer, G. (2000). Was Bernoulli wrong? On intuitions about sample size. *Journal of Behavioral Decision Making, 13,* 133–139.

Sedlmeier, P., & Köhlers, D. (2001). *Wahrscheinlichkeiten im Alltag: Statistik ohne Formeln* [Probabilities in everyday life: Statistics without formulas]. Braunschweig, Germany: Westermann.

Shaugnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465–494). New York: Macmillan.

Shepard, R. N. (1992). The perceptual organization of colors: An adaptation to regularities of the terrestrial world? In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 495–532). New York: Oxford University Press.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky, (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge, England: Cambridge University Press.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research.* Cambridge, England: Cambridge University Press.

Wolfe, C. R. (1995). Information seeking on Bayesian conditional probability problems: A fuzzy-trace theory. *Journal of Behavioral Decision Making, 8,* 85–108.

Appendix

Median and Mean Percentages, and Standard Deviations and Group Sizes, for All Tests in Studies 1a, 1b, and 2

Measure	Liberal scoring				Strict scoring			
	Test 1	Test 2	Test 3	Test 4	Test 1	Test 2	Test 3	Test 4
Study 1a								
Frequency tree								
Median	10	90	90	90	0	80	90	90
M	18	80	77	80	10	73	75	74
SD	23	20	30	26	16	25	32	31
n	15	15	13	5	15	15	13	5
Frequency grid								
Median	10	75	70	100	5	70	60	90
M	16	71	72	70	10	64	62	63
SD	15	23	18	42	14	23	23	45
n	14	14	12	7	14	14	12	7
Rule training								
Median	0	60	30	20	0	35	20	15
M	6	56	48	41	4	41	42	36
SD	11	37	39	38	8	38	43	39
n	20	20	15	10	20	20	15	10
Control								
Median	0		5		0		5	
M	10		18		8		18	
SD	22		28		18		29	
n	5		4		5		4	
Study 1a, complete data sets								
Frequency tree (n = 5)								
Median	10	90	90	90	0	80	90	90
M	12	86	86	80	0	76	86	74
SD	13	9	21	26	0	6	15	31
Frequency grid (n = 7)								
Median	10	90	70	100	0	70	70	90
M	16	79	73	70	10	70	66	63
SD	18	25	22	42	19	22	28	45
Rule training (n = 10)								
Median	0	65	45	20	0	40	40	15
M	4	55	53	41	4	42	46	36
SD	7	33	37	38	7	36	42	39
Study 1b								
Frequency tree								
Median	0	86	79	86	0	71	71	86
M	10	71	67	69	4	61	60	66
SD	14	32	37	37	10	37	36	40
n	28	28	28	28	28	28	28	28
Rule training								
Median	0	64	57	43	0	50	57	43
M	10	55	47	47	5	46	42	40
SD	17	38	39	41	14	38	39	39
n	28	28	28	28	28	28	28	28
Bonus								
Frequency tree								
Median	0	79	71	86	0	57	71	86
M	9	68	58	65	2	58	55	60
SD	14	34	41	39	8	38	38	43
n	14	14	14	14	14	14	14	14
Rule training								
Median	0	71	71	71	0	71	57	57
M	13	63	58	63	8	53	56	56
SD	18	36	37	38	17	36	36	38
n	13	13	13	13	13	13	13	13

(Appendix continues)

Appendix (*continued*)

Measure	Liberal scoring				Strict scoring			
	Test 1	Test 2	Test 3	Test 4	Test 1	Test 2	Test 3	Test 4
				Study 1b (*continued*)				
No bonus								
Frequency tree								
Median	0	86	86	86	0	71	71	86
M	11	75	76	73	6	63	64	71
SD	15	32	31	36	12	37	34	38
n	14	14	14	14	14	14	14	14
Rule training								
Median	0	57	14	14	0	29	0	0
M	7	49	38	34	3	41	30	27
SD	15	40	39	39	11	41	38	35
n	15	15	15	15	15	15	15	15
				Study 2				
Frequency tree								
Median	14	93	93	100	0	86	86	86
M	17	85	85	85	8	76	76	76
SD	26	21	22	24	25	25	30	34
n	24	24	24	21	24	24	24	21
Probability tree								
Median	14	86	79	57	0	86	71	57
M	17	77	73	49	5	73	68	47
SD	15	24	30	41	13	23	34	42
n	24	24	24	21	24	24	24	21
Rule training								
Median	14	86	86	50	0	86	86	21
M	18	70	70	48	9	69	66	41
SD	22	31	35	39	20	32	40	44
n	23	23	23	18	23	23	23	18

Note. The appendix shows median and mean percentages of Bayesian solutions, as well as standard deviations (*SD*) and group sizes (*n*) for all Tests in Studies 1a, 1b, and 2. The results are shown according to both a liberal and a strict scoring criterion (see text). For Study 1a, the data of those participants who took part in all three sessions are shown separately (Study 1a, complete data set). For study 1b, data are also shown separately for those participants who had the chance to earn a monetary bonus ("Bonus") and those who did not ("No bonus"). Values for medians, means, and standard deviations are percentages rounded to the next digit.

Received February 24, 1998
Revision received November 16, 1999
Accepted July 4, 2000 ■