Thinking with Data

The chapters in *Thinking With Data* are based on presentations given at the 33rd Carnegie Symposium on Cognition. The Symposium was motivated by the confluence of three emerging trends: (1) the increasing need for people to think effectively with data at work, at school, and in everyday life, (2) the expanding technologies available to support people as they think with data, and (3) the growing scientific interest in understanding how people think with data.

What is thinking with data? It is the set of cognitive processes used to identify, integrate, and communicate the information present in complex numerical, categorical, and graphical data. This book offers a multidisciplinary presentation of recent research on the topic. Contributors represent a variety of disciplines: cognitive and developmental psychology; math, science, and statistics education; and decision science. The methods applied in various chapters similarly reflect a scientific diversity, including qualitative and quantitative analysis, experimentation and classroom observation, computational modeling, and neuroimaging. Throughout the book, research results are presented in a way that connects with both learning theory and instructional application.

The book is organized in three sections:

- Part I focuses on the concepts of uncertainty and variation and on how people understand these ideas in a variety of contexts.
- Part II focuses on how people work with data to understand its structure and draw conclusions from data either in terms of formal statistical analyses or informal assessments of evidence.
- Part III focuses on how people learn from data and how they use data to make decisions in daily and professional life.

Edited by
Marsha C. Lovett • Priti Shah

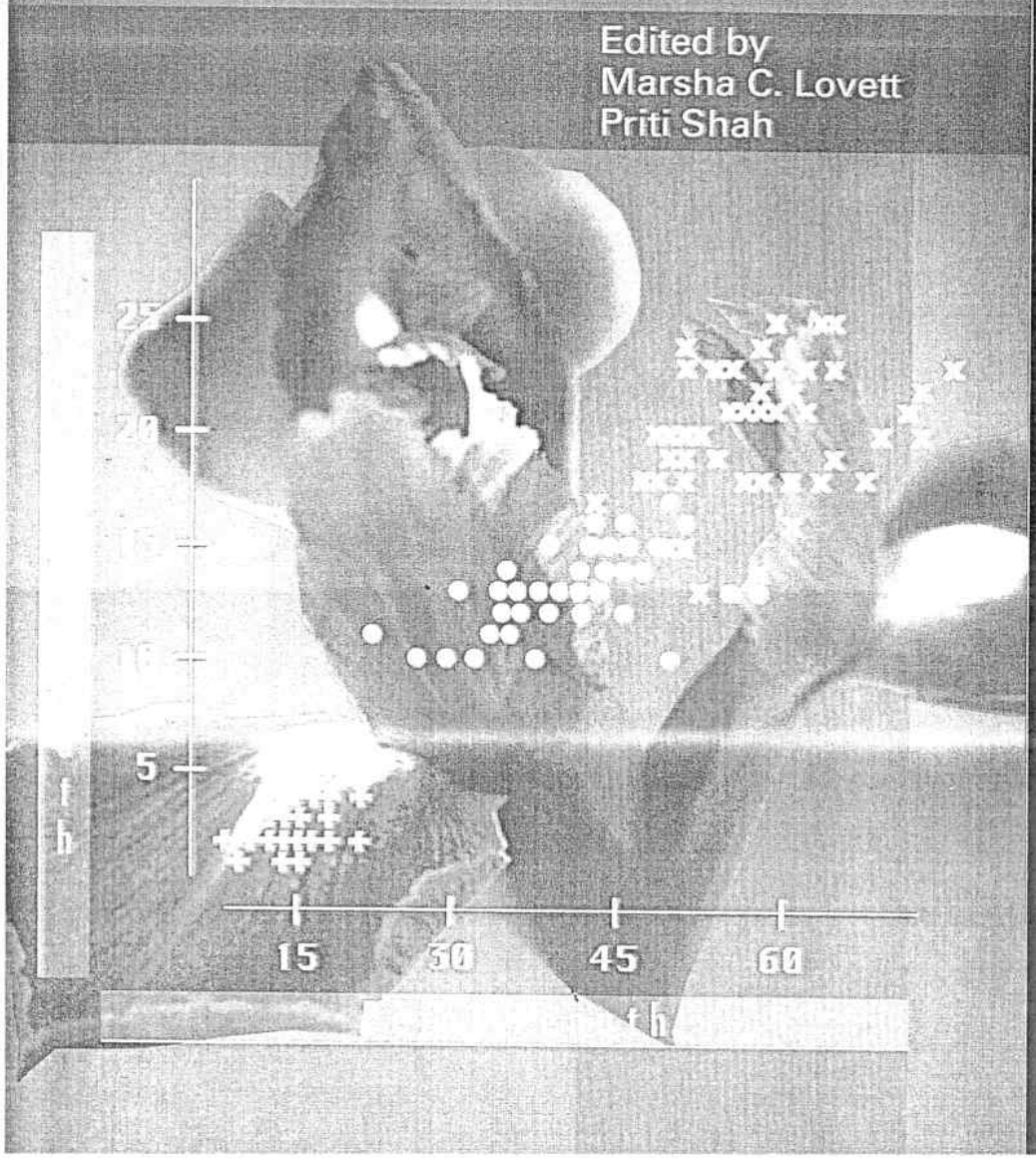
ISBN 0-8058-5422-3

LEA Lawrence Erlbaum Associates
Taylor & Francis Group
www.psypress.com

an informs business

Thinking with Data

Edited by
Marsha C. Lovett
Priti Shah



Wasserman, E. A., Kao, S.-F., Van Hamme, L. J., Katagiri, M., & Young, M. E. (1996). Causation and association. In D. R. Shanks, K. J. Holyoak & D. L. Medin (Eds.), *Causal learning: The psychology of learning and motivation, vol. 34* (pp. 207–264). San Diego: Academic Press.

White, P. A. (2000). Causal judgment from contingency information: Relation between subjective reports and individual tendencies in judgment. *Memory & Cognition, 28*, 415–426.

White, P. A. (2003a). Causal judgment as evaluation of evidence: The use of confirmatory and disconfirmatory information. *The Quarterly Journal of Experimental Psychology, 56A*, 491–513.

White, P. A. (2003b). Effects of wording and stimulus format on the use of contingency information in causal judgment. *Memory & Cognition, 31*, 231–242.

White, P. A. (2003c). Making causal judgments from the proportion of confirming instances: The pci rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29*, 710–727.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford, UK: Oxford University Press.

16

Statistical Reasoning: Valid Intuitions Put To Use

Peter Sedlmeier
Chemnitz University of Technology

Is it true that human minds "are not built (for whatever reason) to work by the rules of probability" as Gould (1992, p. 469) contends? Do we have to "stumble along ill-chosen shortcuts to reach bad conclusions" when we have to reason statistically as McCormick (1987, p. 24) thinks? Until recently, the answer of the majority of researchers in the field of judgment and decision making might have been summarized as a resounding "yes." The main reason for the negative answer has been expressed by Kahneman and Tversky (1973): "In making predictions and judgments under uncertainty, people do not appear to follow the calculus of chance or the statistical theory of prediction. Instead, they rely on a limited number of heuristics, which sometimes yield reasonable judgments and sometimes lead to severe and systematic errors" (p. 237). Such systematic errors or *biases* have indeed been shown again and again in people's judgments (e.g., Kahneman, Slovic, & Tversky, 1982; Piattelli-Palmarini, 1994) and have been accorded the status of *cognitive illusions* (Kahneman & Tversky, 1996). Because of the persistence of these cognitive illusions, there has been little hope for the effectiveness of training programs: "Attempts to train people not to think representatively and not to be influenced by availability or other biases have not been very successful," perhaps because "it is impossible for us to think in a way we do not think" (Dawes, 1988, p. 142). Heuristics can be seen as intuitive reactions that arise when one is asked to solve statistical reasoning problems (see Tversky & Kahneman, 1974, p. 1124). The not very encouraging résumé might then be: When reasoning statistically, we often rely on invalid intuitions that lead to errors and that are immune to training attempts.

Is this really so? My main argument in this chapter is that although invalid intuitions may lead to errors in statistical reasoning, even statistically naïve persons hold valid statistical intuitions that can help them to solve probability problems spontaneously, given that the problems are presented in a suitable representational format. Moreover, these valid intuitions can be of much help in designing efficient training programs to teach different aspects of statistical reasoning. I begin with a broad perspective on intuitive problem solving, to show that valid statistical intuitions are just a special case of a general connection between representational format and intuitive abilities. I then illustrate this idea with two types of problems commonly used in research on statistical reasoning: Bayesian inference and the impact of sample size.[1] Then I propose an associative learning explanation for why and when we have valid statistical intuitions at our disposal. Finally, I discuss the possible uses of these intuitions in training programs.

VALID INTUITIONS IN EVERYDAY PROBLEM SOLVING

A large number of our daily, routine problem-solving activities happen in an intuitive manner (Sedlmeier, 2005). The argument I later apply to statistical reasoning is that a crucial precondition for triggering valid intuitive responses is a suitable external representational format.[2] To clarify this idea and to show how general the phenomenon is, let us look at some examples.

"Everyday Things"

Norman (1988) presented numerous examples from daily life in which the "affordances" of things dramatically change with the way these things are represented. For instance, the way a door handle is connected to a door may evoke the intuitive response to push or to pull or may not evoke any reasonable response at all. Or, consider the light switches for a big lecture hall: in their usual vertical arrangement they do not evoke a clear intuitive response if you want to switch on a certain row of lamps. If the arrangement of light switches instead maps the locations of the lamps, the task of switching on the right lamps can be solved intuitively, as Norman (1988) demonstrated. This principle of mapping, that is, arranging switches or controls according to the spatial locations

[1]The same argument can also be made for two other types of commonly used statistical problems: reasoning about the probability of conjunctive events and about simple conditional probabilities (see Sedlmeier, 2000).

[2]Intuitive responses are "reached with little apparent effort, and typically without conscious awareness. They involve little or no conscious deliberation" (Hogarth, 2001, p. 14).

of the things to be turned on and off, describes a very powerful way of evoking intuitive responses. Further examples are the mapping of stove burners and controls (for standard stoves, there is no satisfactory mapping) and the mapping between faucets and showers (finding the hot water on the first attempt is sometimes not easy). Norman (1993) also calls this the *naturalness principle*: "Experiential cognition is aided when the properties of the representation match the properties of the thing being represented" (p. 72).

Logic

Consider the following two problems (adapted from Cosmides, 1989, p. 192):

> *Problem A:*
> Part of your new clerical job at the local high school is to make sure that student documents have been processed correctly. Your job is to make sure the documents conform to the following alphanumeric rule: "If a person has a 'D' rating, then his documents must be marked code '3'." You suspect the secretary you replaced did not categorize the students' documents correctly. The cards below have information about the documents of four people who are enrolled at this high school. Each card represents one person. One side of a card tells a person's letter rating and the other side of the card tells that person's number code.

Now four cards were shown to the participants, labeled with "D," "F," "3," and "7," and participants are asked to indicate those card(s) one definitely needs to turn over to see if the documents of any of these people violate the above rule. Only 4–25% of college students chose the two correct cards, "D" and "7."[3] However, the proportion of correct solutions rose to 75% in the following problem:

> *Problem B:*
> In its crackdown against drunk drivers, Massachusetts law enforcement officials are revoking liquor licenses left and right. You are a bouncer in a Boston bar, and you'll lose your job unless you enforce the following law. "If a person is drinking beer, then he must be over 20 years old." The cards below have information about four people sitting at a table in your bar. Each card represents one person. One side of a card tells what a person is drinking and the other side of the card tells that person's age. Indicate only those card(s) you definitely need to turn over to see if any of these people are breaking this law.

[3]This is a version of the famous "Wason selection task." The logical conclusion "If *P* then *Q*" can be falsified if *P* is true but *Q* is not. The four cards in the problem represent *P* (D), *not-P* (F), *Q* (3), and not-Q (7). There are two ways in which the rule can be falsified: either it turns out that on the other side of the D card there is a 7 (*not-Q*), or that on the other side of the 7, there is a D (*P*).

The cards presented to the participants now read "drinking beer," "drinking coke," "25 years old," and "16 years old." It is almost immediately clear to most people that one has to check whether that person drinking beer is over age, and whether the 16-year-old is drinking beer. Although the two problems are both variants of the same type of logical problem, the content makes a big difference. The first situation is very unfamiliar and does not evoke intuitive responses whereas one does not need logical conclusions to arrive at the logical solution in the second problem: it comes intuitively.

School Math

Fischbein and his collaborators examined mathematical intuitions in young students. One set of problems they used consisted of the following two questions (Fischbein, 1994, p. 234):

1. From 1 quintal of wheat, you get 0.75 quintals of flour. How much flour do you get from 15 quintals of wheat?
2. 1 kilo of a detergent is used in making 15 kilos of soap. How much soap can be made of 0.75 kilos of detergent?

These two problems have the same solution: one just has to multiply two numbers, 15 and 0.75. However, the solution rates across fifth-, seventh-, and ninth-grade students, although they did not differ so much across age groups, were quite different for these two problems: about 75% correct solutions for the first problem and about 25% for the second. Why? Fischbein (1994) explains the difference with students applying the intuition that "multiplication makes bigger." If the second number—the number by which the first is multiplied—is larger than 1, this means that the result of the multiplication is larger than the first number: multiplication makes bigger. This intuition can be applied to the first problem but not the second. Note that the crucial difference between these two problems seems to be just the order in which information is presented.

General Principle

There are many other areas where one can find similar results. Here is one more example: Many studies in South America have shown that children and adolescents, as well as carpenters and fishermen with almost no schooling, are successfully able to sell fruit, build houses, and charge correct fish prices in an intuitive way—with *street math*. However, this ability breaks down if the same persons are given the identical tasks formulated in versions used in school training (Nunes, Schliemann, & Carraher, 1993). What do all these examples show? Although they are quite different, the general principle is the

same: The external representational format can make a big difference. The difference may consist in spatial arrangement, matters of content or context, or even the order of numbers—if the external representation matches a valid intuition, the corresponding decisions are made quickly (and correctly) and problems can be solved easily.

VALID STATISTICAL INTUITIONS

The idea exemplified in the last paragraph—that the right representational format may help in intuitive problem solving—will now be applied to statistical reasoning. For illustration, I have chosen two classes of problems that are often used in research on statistical reasoning: Bayesian problems and sample-size problems. For each of the two classes of problems, I first present the (standard) difficult version and then show how slight changes in representational format make them easier. Then I describe the valid intuition that makes the problems easier, in each case. The intuition that I postulate to work in the case of Bayesian problems I call *ratio intuition*; and the intuition that is assumed to work in the case of sample-size tasks is the *size-confidence intuition*.

Bayesian Problems, Difficult Version

In Bayesian problems the probability of an event is revised in the light of new information. Here is an example, the *mammography problem*, originally examined by Casscells, Schoenberger, and Graboys (1978), in an adapted version (part headings inserted for discussion only):

> *Introductory text*
> 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:
> *Base rate (prior probability)*
> The probability that a randomly chosen woman who undergoes a mammography exam will have breast cancer is 1%.
> *Hit rate*
> If a woman undergoing a mammography exam has breast cancer, the probability that she will test positive is 80%.
> *False alarm rate*
> If a woman undergoing a mammography exam does not have breast cancer, the probability that she will test positive is 10%.

Question
What is the probability that a woman who has undergone a mammography exam actually has breast cancer, if she tests positive?

The event in question is that a randomly chosen woman has breast cancer. Without any further knowledge, the (prior) probability for that to be true is 1% or .01—the base-rate information given in the problem. The new information is that the mammography test was positive. This information is connected to two other pieces of information: the hit rate, p(positive test|cancer), and the false alarm rate, p(positive test|no cancer), which are given in the problem as 80% and 10%, respectively. How should the prior probability be revised to arrive at the sought for posterior probability, p(cancer|positive test)? Bayes' theorem provides the solution: equation 1.

The available evidence indicates that Bayesian problems of this and other types are very difficult for laypeople as well as professionals, with correct solution rates of usually less than 10% (e.g., Casscells et al., 1978; Eddy, 1982; Kahneman & Tversky, 1972; Koehler, 1996). The main problem seems to be that often people do not take the base rate into account, a neglect that has been termed the base-rate fallacy. The conclusion is clear: "The genuineness, the robustness, and the generality of the base-rate fallacy are matters of established fact" (Bar-Hillel, 1980, p. 215).

$$p(cancer|pos.) = \frac{p(cancer)\,p(pos.|cancer)}{p(cancer)\,p(pos.|cancer) + p(no\ cancer)\,p(pos.|no\ cancer)} \qquad (1)$$
$$= \frac{.01 \cdot .8}{.01 \cdot .8 + .99 \cdot .1}$$
$$= .075$$

Bayesian Problems Made Easier

How can Bayesian problems be made easier? Gigerenzer and Hoffrage (1995) argued that the format in which the problem information is presented is of crucial importance. They reasoned that people can deal much more easily with natural frequencies, that is, frequencies not normalized with respect to base rates, than with probabilities.[4] Restated in terms of natural frequencies, the mammography problem reads (introductory text remains identical):

[4]Absolute frequencies are not eo ipso natural frequencies. For instance, the information in the mammography problem could also be stated in normalized absolute frequencies: a base rate of 1 in 100, a hit rate of 80 in 100, and a false alarm rate of 10 in 100. These frequencies, 1, 80, and 10, no longer carry information about the relevant base rates.

Base rate (prior probability)
Ten of every 1,000 women who undergo a mammography exam have breast cancer
Hit rate
Eight of every 10 women with breast cancer who undergo a mammography exam will test positive.
False alarm rate
Ninety-nine of every 990 women without breast cancer who undergo a mammography exam will test positive.
Question
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?

Note that not only has the representational format been changed from probabilities (expressed in percentages) to frequencies, also the calculation has become easier: equation 2.

Gigerenzer and Hoffrage (1995) found an increase of Bayesian solutions from 16% with problems formulated in the probability format to 46% when formulated in the natural frequency format. Similar results were found with laypeople (Christensen-Szalanski & Beach, 1982; Cosmides & Tooby, 1996) and experts (Hoffrage & Gigerenzer, 1998; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000).

$$p(cancer|pos.) = \frac{\#(pos. \cap cancer)}{\#\,pos.} \qquad (2)$$
$$= \frac{8}{107}$$
$$= .075$$

The Ratio Intuition

How can intuition help to solve Bayesian problems? My suggestion is that several kinds of probability problems can be solved intuitively by building a ratio between a smaller and a larger number. I call this intuition the *ratio intuition*. The ratio intuition might be seen as an extension of the *intuition of relative frequency*, for which Fischbein (1975) cites ample evidence. I further argue that this intuition only works if the numbers in question represent natural frequencies (see Footnote 4).

What kinds of probability problems am I talking about? Let us consider the situation in Figure 16.1, which depicts the solution to the mammography

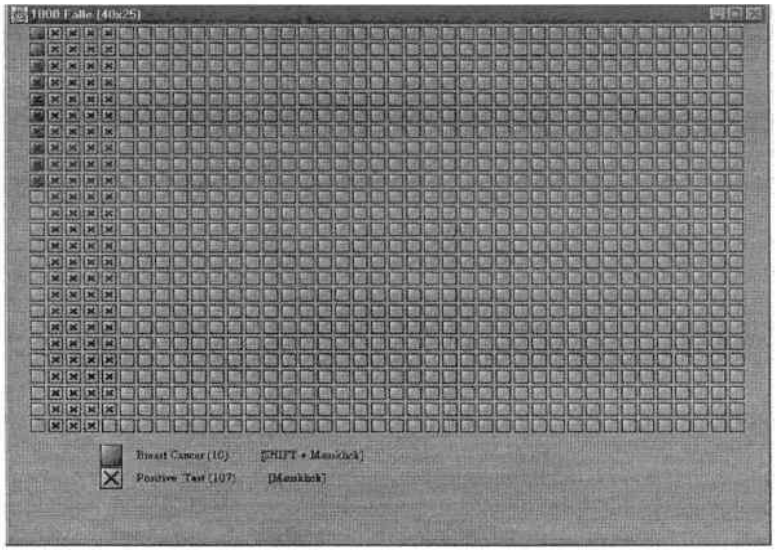


Figure 16.1 Frequency grid representation of the mammography problem. The 10 shaded squares represent the number of women out of 1,000 who can be expected to have breast cancer according to the information given in the mammography problem, and the 107 squares marked by a cross represent the number of women who can be expected to obtain a positive test result. The frequency grid was generated by using the software that accompanies a textbook on elementary probability theory (Sedlmeier & Köhlers, 2001). See text for further explanation.

problem as given by a computerized tutorial that covers the elements of basic probability theory as taught in German high schools (Sedlmeier & Köhlers, 2001).[5] The 1,000 cases (German: *Fälle*) in the problem are represented by squares. Specific squares can be shaded in the program by pressing the Shift key and clicking with the mouse on the square (German: *Mausklick*). In this way 10 squares have been shaded, representing the 1% of the women who have breast

[5]This program is a modification and extension of several more specific programs tested in prior experiments (Sedlmeier, 1999; Sedlmeier & Gigerenzer, 2001; see also the section on "Training Programs in High School"). The program relies heavily on flexible frequency representations, learning by doing, and immediate feedback, corresponding to three of Lovett and Greenhouse's (2000) five principles of learning (Principles 1, 3, and 5).

cancer. Out of these 10, the 8 (80% of 10) with a positive test result are marked with a cross (by clicking on the square with the mouse without pressing any other key), and so are a further 99 squares, representing the 10% of women with a false-positive test (out of 990). The ratio that has to be built to solve the mammography problem is 8/107, that is, one has to divide the 8 women who have cancer and get positive test results by all women who get positive test results (see Equation 2). This is an example of how the ratio intuition can be applied to Bayesian tasks. But the special thing with frequency representations is that the solution of Bayesian tasks does not differ from the solution of apparently simpler tasks such as finding the probability of conjunctive events or finding simple conditional probabilities. What, for instance, is the probability that a randomly chosen woman both suffers from breast cancer and gets a positive test result—the probability of a conjunctive event? Figure 16.1 would quickly give the answer: 8/10 or 80%. And what is the conditional probability that a given woman has a positive test result if she does have cancer? We already know this probability—the hit rate from the problem—but if we did not, Figure 16.1 would quickly give the answer: 8/10 or 80%.

So what we see here is that if natural frequencies are used as the representational format, probability problems that are expressed differently in the usual language of probability theory are solved in the same way.[6] The solution in all three cases above is to build a ratio that relates a smaller number to a larger number or, in other words, to calculate a relative frequency. The three cases just differ in the selection of the right reference class: all women in the sample for the conjunctive-probability problem, all women with cancer in the simple conditional-probability problem, and all women with a positive test result in the original Bayesian problem. In many cases, finding the right relative frequency does not seem to be difficult if events are represented in terms of natural frequencies: Even young children show a remarkable sensitivity in both choice and estimation tasks (Huber, 1993; Inhelder & Piaget, 1959/1964; Kuzmak & Gelman, 1986; Reyna & Brainerd, 1994).

However, there is no guarantee that the right reference class will be chosen automatically: Especially when people seek to confirm hypotheses they are already holding and if they are not restricted in the way they sample information they may tend to use biased reference classes (Fiedler, Brinkmann, Betsch, & Wild, 2000; Klayman & Ha, 1989). In addition, the

[6]The usual way to arrive at a Bayesian solution—p(cancer|positive result) in our example—is to use Bayes' theorem (Equation 1). The probability of the conjunctive event in our example is usually calculated as p(cancer & positive result) = p(cancer|positive result)p(positive result) = p(positive result|cancer)p(cancer), and the simple conditional probability would be calculated as p(positive result|cancer) = p(cancer & positive result)/p(cancer).

selection of a reference class is also influenced by whether people perceive base rates to be relevant and reliable (Ginossar & Trope, 1980; Koehler, 1996). If the wrong base rates are chosen, the ratio intuition would, of course, yield biased results.

Let us now turn to another type of problem that examines how well the impact of sample size on confidence judgments about proportions and means is understood.

Impact of Sample Size: Problematic Results

After several studies on the impact of sample size on statistical reasoning, Kahneman and Tversky (1972) concluded that "a strong tendency to underestimate the impact of sample size lingers on despite knowledge of the correct rule and extensive statistical training" (p. 445). One of the problems they used was the following, the *maternity ward* problem (part headings inserted for discussion only):

Introductory text

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

Specification part

For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?

Answer alternatives

A. The larger hospital

B. The smaller hospital

C. About the same

What is the correct solution for this task? The question refers to the difference in the variances of empirical sampling distributions or, to be more exact, to the difference between the number of proportions that are expected to fall beyond a value of 60% in two sampling distributions, one for the larger and one for the smaller hospital. These distributions are binomial distributions with $p = .5$ (a 50% chance that a baby born is a boy) and sample sizes of $n = 45$ (larger hospital) and $n = 15$ (smaller hospital), respectively. The correct answer can be inferred from the result of a simulation shown in Figure 16.2.

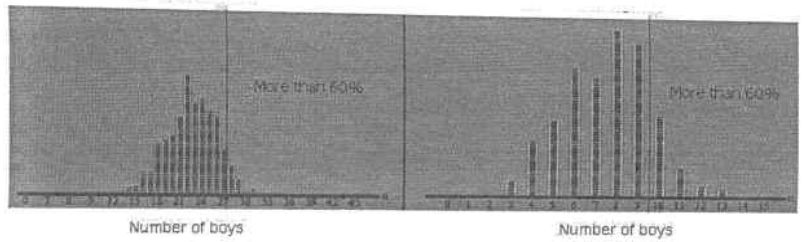


Figure 16.2 Illustration of the solution of the maternity ward task (sampling distribution version). The simulation results (relying on Bernoulli trials with $p = .5$) were obtained by using the software that accompanies a textbook on elementary probability theory (Sedlmeier & Köhlers, 2001).

The figure shows two empirical sampling distributions, consisting of 365 proportions of "baby boys" each (corresponding to the 365 days of the year). The results over a year are shown for the smaller hospital with proportions calculated from 15 "births" each (Fig. 16.2, left) and the larger hospital with proportions calculated from 45 "births" each (Fig. 16.2, right). The result in Figure 16.2 is typical: because sampling distributions with larger sample sizes can be expected to be more closely centered around the expected value (50%), the chances for proportions to deviate from the expected value beyond a given noncentral value are higher in the smaller hospital.

Participants given this and similar tasks usually (and erroneously) prefer C, that is, the option that says "no difference"; and the percentages of correct solutions are only around 20% (e.g., Kahneman & Tversky, 1972; Murray, Iding, Farris, & Revlin, 1987; Swieringa, Gibbins, Larsson, & Sweeney, 1976). These results are echoed in a summarizing statement by Reagan (1989, p. 57): "The lesson from 'sample size research' is that people are poorly disposed to appreciate the effect of sample size on sample statistics."

Impact of Sample Size Re-Examined

Already in the 1950s, Piaget and Inhelder (1951/1975) found children from the age of 11 or 12 to be sensitive to the impact of sample size on the quality of estimates. Should adults have lost this sensitivity? There is indeed evidence for this suspicion (see the previous paragraph) but could the decline be so dramatic? Apparently not—otherwise one would not expect statements like

this one: "Overall, subjects did quite well as intuitive statisticians in that their judgments tended, over the experiments as a whole, to move in the direction required by statistical theory as the levels of Mean Difference, Sample size and Variability were varied" (Evans & Pollard, 1985, pp. 68–69). Gerd Gigerenzer and I took a closer look at the sample size literature and found a huge variation in solution rates (Sedlmeier & Gigerenzer, 1997). After considering several possibilities, the explanation most plausible to us was that two different types of tasks have been used in the pertinent research: one type, of which the original maternity ward task is an example, we termed a "sampling-distribution task." In 29 studies in which participants had to solve such tasks the median solution rate was 33%, a figure that can be expected by chance (if one decides randomly among the three alternatives). However, in a slightly different type of task, the median solution rate was dramatically higher: 76%. We termed this type of task a "frequency-distribution task." Here is an example, a variant of the maternity ward problem:

Introductory text

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

Specification part

Which hospital do you think is more likely to find on a given day that more than 60% of the babies born were boys?

Answer alternatives

A. The larger hospital

B. The smaller hospital

C. About the same

Note that this task differs from the original one only in the "Specification part" but this difference seems to be a crucial one. Why? Because in this version of the task, the question is about two distributions of raw scores or "frequency distributions," the distributions of baby boys and baby girls in the two hospitals on a given day. The proportion from a single sample in the larger hospital is compared to the proportion from a single sample in the smaller hospital.

The Size-Confidence Intuition

Why is the frequency-distribution version of the maternity ward problem so much easier than the corresponding sampling-distribution problem?

Centuries ago Jacob Bernoulli argued that "even the stupidest man knows by some instinct of nature per se and by no previous instruction" that the greater the number of confirming observations, the surer the conjecture (Gigerenzer et al., 1989, p. 29). This is an early formulation of what Gerd Gigerenzer and I called the *size-confidence intuition* (Sedlmeier & Gigerenzer, 1997). The size-confidence intuition conforms to the *empirical law of large numbers* and is not a mathematical law, but it can be experienced when drawing random samples: As samples become larger, means or proportions calculated from these samples tend to become more accurate estimates of population means or proportions. This tendency may at times not hold, that is, larger samples may lead to more inaccurate results, but the law holds on average (Sedlmeier & Gigerenzer, 2000).

The size-confidence intuition makes it easy to solve frequency-distribution tasks. For instance, in the frequency-distribution version of the maternity ward task, the result from the larger hospital can be expected to be closer to the "true" 50% than the result from the smaller hospital, or—in other words—a large deviation of the proportion of baby boys from 50% is more likely in the smaller hospital. However, the size-confidence intuition is not directly applicable to sampling-distribution tasks, which explains the difference in solution rates found in the relevant studies (Sedlmeier, 1998; Sedlmeier & Gigerenzer, 1997).

One could argue that what we called sampling-distribution tasks can be solved by repeatedly applying the empirical law of large numbers. This is immediately evident to persons trained in statistics but the empirical evidence indicates that laypeople are not able to do so spontaneously. So the crucial difference for the size-confidence intuition is whether frequency-distribution tasks or seemingly similar sampling-distribution tasks are to be solved. In addition, the spontaneous use of this intuition seems to profit largely from what could be called the *dynamical frequency format*. Samples naturally come as frequencies of events but it seems to make a difference whether the sampling process is only described verbally or whether one can directly experience it. When participants had the chance to perform simulations of the sampling process with the help of a computer program, the rates of correct solutions were about 25% higher than when problems were just described to them in texts (Sedlmeier, 1998). This finding indicates that frequency representation per se does not always help automatically—the match between visual features and the perceived meaning of these features is decisive (see also Shah & Hoeffner, 2002).

There are several other explanations for why such a huge variance in the solution rates for sample-size tasks can be found in the literature. The factors responsible include the ratio between sample sizes (Murray et al., 1987), the extremity of cut-off percentages (e.g., "more than 80%" instead of "more than

60%," Bar-Hillel, 1982), the part of the distribution to which the question refers (Well, Pollatsek, & Boyce, 1990), the complexity of the problems (Evans & Dusoir, 1977), and the salience of chance factors (Nisbett, Krantz, Jepson, & Kunda, 1983). However, none of the examined factors was found to have an influence on solution rates nearly as large as the distinction between frequency- and sampling-distribution tasks suggested by the size-confidence intuition (for details see Sedlmeier, 1998; 2006).

STATISTICAL INTUITIONS AS THE RESULT OF ASSOCIATIVE LEARNING

I have just postulated the existence of two intuitions that might be helpful in solving statistical reasoning problems. Here I suggest how these intuitions might arise: They are the result of associative learning. The specific associative learning model I am proposing is the PASS (probability associator) model, which was developed to simulate probability and relative frequency judgments (Sedlmeier, 1999, 2002). I first describe how PASS works in principle and then discuss how it may account for the ratio and the size-confidence intuitions.

PASS: Learning And Representation

PASS encodes events (hereafter, "event" stands for "events," "objects," "persons," etc.) consecutively. These events may exist either in reality or in the imagination. An event is represented by its features, to which the associative learning mechanism applies. Figure 16.3 shows a simplified representation of the events needed to model the mammography problem. PASS works with distributed representations, that is, "breast cancer," "positive test," and "female" would be represented by patterns of features. For the sake of simplicity, these patterns have been compressed here to one feature each. If the feature is present, this is shown as a filled circle in Figure 16.3; if the feature is absent, the circle is empty. For instance, "Ms. A," one of the 9 women represented in Figure 16.3, does not have breast cancer, nor did she obtain a positive test result, whereas "Ms. B's" breast cancer has been correctly diagnosed.

The core of PASS is a neural network that encodes features by its input nodes, modifies the associations between features by changing the weights between nodes, and elicits reactions by producing activations at its output nodes. Learning consists in modifying the associations between the features in memory (in this simplified example, the memory consists of only three features and their associations). PASS operates in discrete time steps. At each time step, the features of the event towards which the attention is directed are encoded. Learning takes place at every time step. If features co-occur in an

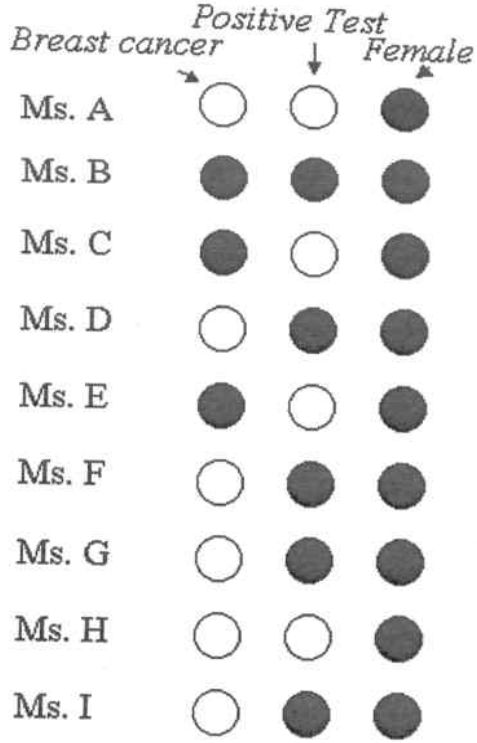


Figure 16.3 A simplified representation of part of the information from the mammography problem as used by the PASS model. Only three features ("breast cancer," "positive test," and "female") are shown per "event." Filled circles indicate that the respective feature is present, whereas empty circles indicate the absence of that feature. For instance, "Ms. A." is female but does not have breast cancer, nor has she tested positive.

event—such as when PASS encounters the featural description of "Ms. B." in Figure 16.3, the association between them is strengthened; if this is not the case, the association between the respective features becomes weaker—such as the association between "breast cancer" and "positive test" when PASS encounters the description of "Ms. A." or that of "Ms. C."[7] This learning

[7]PASS would use two different "forgetting rules" "for the two cases, a rule that simulates memory decay in the first case, and interference in the second.

process results in a memory matrix that is updated after every time step and that contains the association strengths between all the features in memory. When PASS is prompted with an event, that is, the features that define this event, activations occur at all output nodes (activations range from 0 to 1). The sum of these activations is taken as PASS's response. The more often features occur together, the higher is PASS's response when prompted with an event that is described by these features or at least a large number of them.

PASS's Account of the Ratio Intuition

Estimating relative frequencies and thus arriving at ratio estimates is PASS's basic task. For its estimates, PASS essentially follows the definition of relative frequency. It first looks for the conjunctive events—for example, the joint occurrences of the features "breast cancer" and "positive test" in the mammography problem—and determines their summed activation (in Fig. 16.3, only the event "Ms. B." would be treated in this way). Then it sums the activations for all events of the reference class—for example, all occurrences of the feature "positive test" in the mammography problem (5 of the 9 events in Fig. 16.3 would count as this). And finally it divides the first sum of activations by the second and takes this as an estimate of the relative frequency or ratio (see Sedlmeier, 1999, 2002; for a similar model, see Dougherty, Gettys, & Ogden, 1999). This is how PASS would solve the mammography problem. There is ample evidence that building ratio estimates can be done very quickly and is often more accurate when done intuitively (Sedlmeier & Betsch, 2002).

PASS's Account of the Size-Confidence Intuition

The explanation of the size-confidence intuition in the PASS model follows as an emergent property of the way the model learns. In PASS it is assumed that confidence judgments covary with how well the model can differentiate between different events. PASS uses the variance across the output units' activations as an index for confidence: In the beginning of the learning process, there is little variation among the activations of PASS's output units, and PASS would not be very confident in its estimates. The more variation there is in these activations after being prompted with a given event, the better PASS "knows" this event. It can be shown that the variance tends to increase with increasing sample size, thus exhibiting the size-confidence intuition (Sedlmeier, 1998; 1999). Note that there is no additional mechanism involved: The variation of output activations is just a by-product of the learning process.

Why Do We Need Training Despite Valid Intuitions?

One might now ask why training is needed when statistical problems can already be solved intuitively. According to PASS (or similar associative learning models), a problem is solved intuitively if there is a match between the way it processes information and the way the problem information is perceived. In both the mammography problem and the maternity ward problem, information about the events is not presented serially. Can the PASS model be applied to these problems at all? In the case of pictorial representations such as that in Fig. 16.1, one can argue that the squares or other symbolic representations could be nevertheless encoded serially. But what if problems are only described in a text? It seems that the crucial point for whether the PASS model applies (i.e., whether a problem can, in principle, be solved intuitively) is whether the information is transformed in a way that enables the imagination of discrete events (for recent evidence that memories develop with imagination alone, see Mazzoni & Memon, 2003). The importance of imagination in judgmental processes has been stressed repeatedly (e.g., Hogarth, 2001; Kahneman & Tversky, 1982a). So whenever one can expect that texts elicit an imagining of discrete events, either by giving prompts or as the result of specific training, one can expect intuitive judgments according to the PASS model. Let us return to the question just discussed: Training is needed to help "translate" a problem into a suitable representational format. The existence of a valid intuition alone does not automatically lead to its application: it has to be triggered first. Training can also help to increase the likelihood that valid intuitions will be triggered appropriately.

VALID INTUITIONS IN TRAINING

If the serial encoding or imagining of events is a crucial precondition for the use of statistical intuitions, as postulated by the PASS model, how should training programs be designed to make these intuitions work? The most plausible answer seems to be to make trainees learn to translate an unhelpful representational format into a format that can be expected to evoke intuitive and correct solutions. This is what we tried to do. Here, I give a short overview about the empirical evidence with Bayesian training and with training that taught the understanding of the impact of sample size.

Bayesian Training

In several studies, we tested training programs that taught participants to translate probability information into natural frequencies (Sedlmeier, 1997; 1999;

Sedlmeier & Gigerenzer, 2001). The computerized training programs consisted of two parts. The first part led trainees through the solution of two problems, step by step; in the second part, trainees could solve Bayesian problems on their own but the program gave corrections and hints and ensured that every problem could be solved correctly. Among the representational formats used were the frequency grid (Fig. 16.1) and the frequency tree (Fig. 16.4, left; *Baum* is German for "tree"). In the frequency tree training, the number of women in the reference class (here: 1,000) is divided up according to the base rate (here: 1%, that is, 10 out of 1,000 women with breast cancer—see middle part of tree) and then according to the hit rate (here: 80%, that is, 8 out of 10 women), and the false-alarm rate (here: 10%, that is, 99 out of 990 women). The results of these training programs were compared with those obtained in training programs that used the conventional probability format. In the latter, trainees were taught either to fill in Bayes' formula (Equation 1) with the appropriate values (not shown) or to fill in the right values on a probability tree (Fig. 16.4, right). Filling in the values in the probability tree works similar to the frequency tree—the only difference being that the decimal point for all numbers is moved three digits to the left (e.g., 8 becomes .008). In both the frequency and the probability tree the correct solution for p(cancer|positive result) is obtained by dividing the number in the leftmost node by the sum of this number and the number in the third node at the bottom of the tree [e.g., .008 / (.008 + .099) in the case of the probability tree]. The training regimens were held as similar as possible otherwise.

Trainees were tested immediately before and after the training, which (without the testing) took less than an hour on average. In addition, trainees were tested again 1 week and up to 3 months after the training. Figure 16.5 shows a representative result. In this study, participants were quite good with the probability formats immediately after training, but the results after 3 months show the long-term effect of using frequency formats to be clearly superior. Our results were obtained with university students, both in the United States and in Germany, as participants, but similar results were also found with high school students (Wassner, Martignon, & Sedlmeier, 2002). Comparable training regimens that did not use computerized tutorials also yielded similar results, although the absolute success rates were somewhat lower (Ruscio, 2003).

My explanation for the difference in results shown in Figure 16.5 is that the ratio intuition was more likely to be triggered by the frequency-tree representation than by the two probability representations. Is there not a much simpler explanation for this difference: Could it be that frequency representations require a lighter information-processing load than probability representations (compare Equations 1 and 2)? If this explanation holds, one should expect similar results for the frequency tree and the probability tree (but not

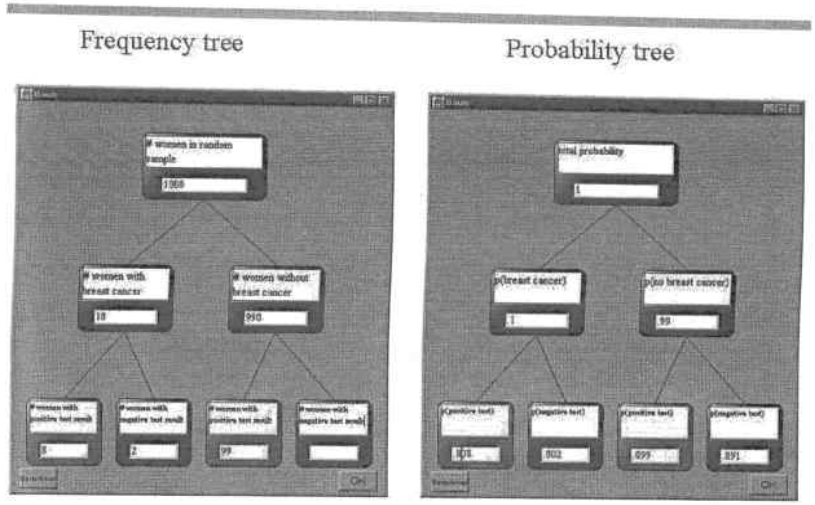


Figure 16.4 Frequency tree and probability tree, similar to those used in the original training studies. The probability tree carries the same information as the frequency tree; only the decimal point is moved three positions to the left. Screen shots are from the software that accompanies a textbook on elementary probability theory (Sedlmeier & Köhlers, 2001). The button at the lower left (*Berechnen* is German for "calculate") can be used to check whether the numbers are inserted correctly.

in the rule training) because in both cases, the simpler Equation 2 or its equivalent for probabilities can be used. However, Figure 16.5 indicates that it is not the difference in information-processing load but the difference in representational format that is decisive.

There have also been other attempts at training Bayesian reasoning, including corrective feedback (e.g., Lindeman, van den Brink, & Hoogstraten, 1988) and directing participants' attention to relevant information (e.g. Fischhoff & Bar-Hillel, 1984; Wolfe, 1995). However, the training effects in these studies were quite modest, overall (see Sedlmeier, 1999).

Training to Solve Difficult Sample-Size Problems

Apparently, there is not much need to train the understanding of frequency-distribution tasks because the size-confidence intuition yields high spontaneous solution rates. But the size-confidence intuition can be put to use to solve difficult sampling-distribution tasks by making the connection between

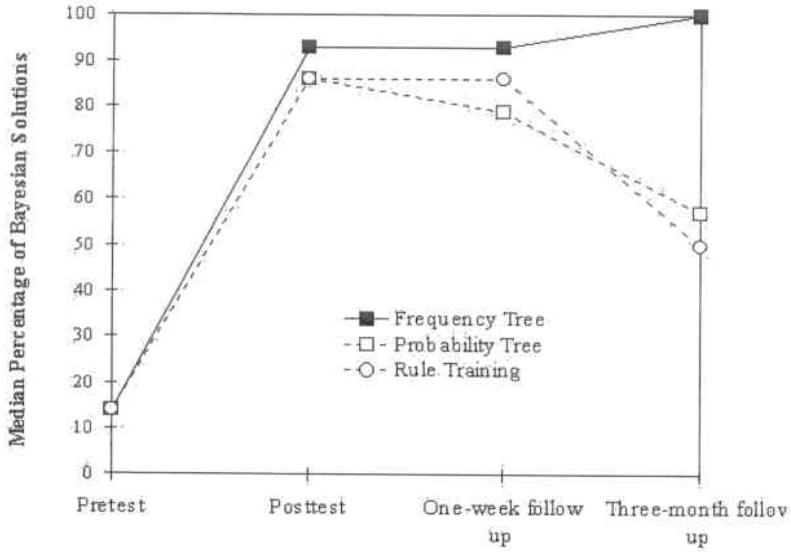


Figure 16.5. A typical result in training studies on Bayesian reasoning, adapted from Sedlmeier (1999).

frequency and sampling distributions clear. Figure 16.6 shows how this was accomplished in the training studies (see Sedlmeier, 1999, chaps. 9–11). The figure shows screen shots collected during different stages of th
e computerized tutorial (Sedlmeier & Köhlers, 2001). A dynamical frequency format was realized with the help of a virtual urn that contains a population distribution (e.g., a distribution of 50% boys and 50% girls, in the case of the maternity ward problem). The sampling process was simulated by letting a black bar with a funnel end stir the contents of the urn with random movements. After a while, the stick turned yellow and became "magnetic," thus attracting the nearest "event" (a boy or a girl, in the case of the maternity ward problem) and pulling it out (see Fig. 16.6, top). This elementary sampling process was repeated until a sample had the planned size. Figure 16.6 (middle) shows three samples of Size 15, depicting frequency distributions of boy and girl births in the smaller hospital on three arbitrary days. The proportions calculated from these samples were then put on the frame for an empirical sampling distribution—"day" by "day" in the case of the maternity ward problem. The collection of proportions finally resulted in empirical sampling distributions as

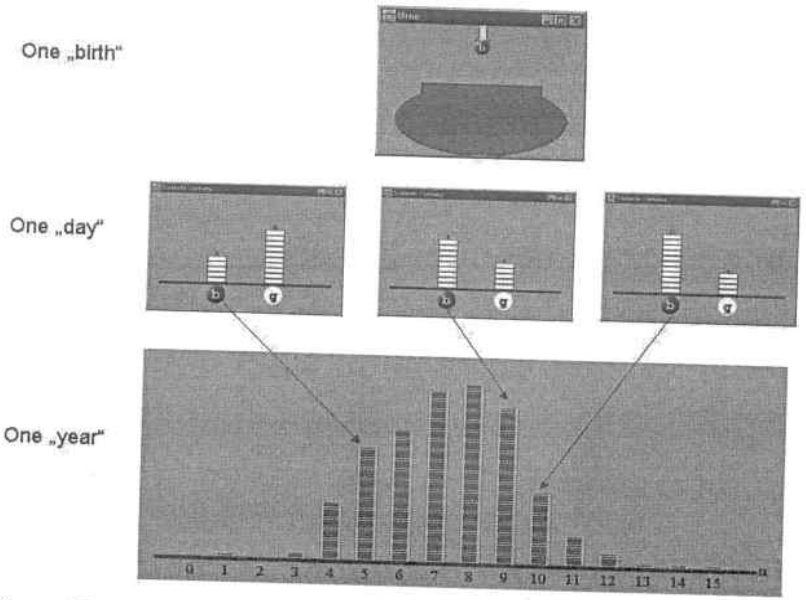


Figure 16.6 Steps ranging from sampling single results to eventually obtaining sampling distributions, using the information from the maternity ward problem. The top of the figure shows an urn that contains a population distribution of all potential births (50% boys and 50% girls). "Births" that result in either "boys" (b) or "girls" (g) are drawn one by one. The middle part shows three arbitrary days in the smaller hospital. The respective proportions of baby boys are placed onto an empirical sampling distribution (lower part) that depicts the proportions of baby boys over 365 days. Screen shots are from the software that accompanies a textbook on elementary probability theory (Sedlmeier & Köhlers, 2001).

specified in the respective problems (e.g., sampling distributions consisting of 365 days, in the case of the maternity ward problem—Fig. 16.6, bottom). Trainees could watch this process concurrently for samples of different sizes (e.g., for the small and the large hospital, see Fig. 16.2).

Similar to the Bayesian training studies, tests were administered immediately before and after the training, as well as 1 week and 5 weeks afterwards. The problems used were of the maternity-ward type (original version) and of another type in which participants had to construct sampling distributions themselves. Here is an example of the latter (adapted from Kahneman & Tversky, 1972, p. 437):

Imagine that in a certain country demographic properties in different regions are recorded. In one region (Region A) there are about 10 births per day and in another region (Region B) there are about 40 per day. Every day, the proportion of boys and girls is registered.
Please estimate the percentages of female births that can be expected in both regions over a period of 100 days. Just divide the 100 days over the categories, for each region.

Region A	*Region B*
About 10 births daily	About 40 births daily
__ Up to 5% girls	__ Up to 5% girls
__ 6% to 15% girls	__ 6% to 15% girls
__ 16% to 25% girls	__ 16% to 25% girls
__ 26% to 35% girls	__ 26% to 35% girls
__ 36% to 45% girls	__ 36% to 45% girls
__ 46% to 55% girls	__ 46% to 55% girls
__ 56% to 65% girls	__ 56% to 65% girls
__ 66% to 75% girls	__ 66% to 75% girls
__ 76% to 85% girls	__ 76% to 85% girls
__ 86% to 95% girls	__ 86% to 95% girls
__ 96% to 100% girls	__ 96% to 100% girls.

The solution rates in this type of problems are even lower than those in maternity-ward-type tasks: Usually participants' sampling distributions do not differ at all (Fischhoff, Slovic, & Lichtenstein, 1979; Kahneman & Tversky, 1972; Olson, 1976; Sedlmeier, 1992; Teigen, 1974).

Figure 16.7 shows the results for such construction tasks before and after a computerized sample-size training (Sedlmeier, 1999). Results were scored as correct if the number filled in for the middle part was larger for the larger sample and if the numbers for the extreme parts of the sampling distributions were larger for the smaller sample, and if no anomalies (e.g., bimodality) were observed. Apparently the short training (less than half an hour) was quite effective. Although the new problems presented in the testing sessions were not solved as well as the problems used in training, stable solution rates of 85% were obtained, a figure that has not been observed in the literature, so far.

To the best of my knowledge there has been only one other attempt to develop sample-size training using sampling distribution tasks (not termed so, though). This training (Well et al., 1990, Study 4) consisted of oral explanations, a demonstration of random sampling using paper slips, and a simulation on a computer screen. However, only 24% of the participants solved a sample-size task correctly afterward. Other training attempts used frequency-distribution tasks (e.g., Fong, Krantz, & Nisbett, 1986) and taught participants the rule that sample parameters approach population parameters as a function of sample size and

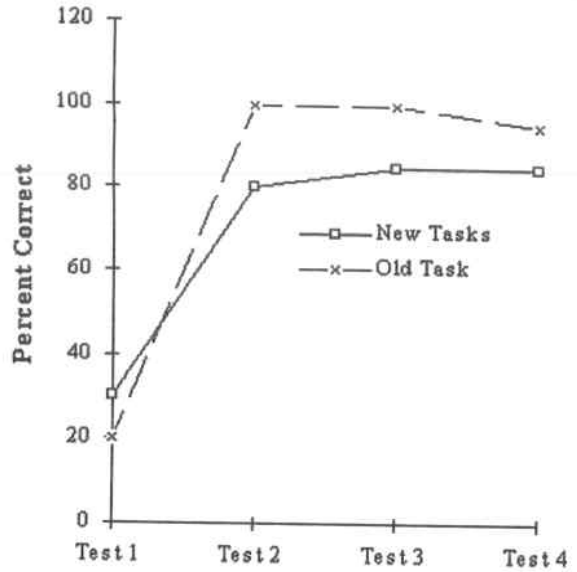


Figure 16.7 Solution rates obtained in tests on sampling-distribution problems (construction tasks) immediately before and after a sample-size training (Test 1 and Test 2), as well as 1 and 5 weeks afterwards (Test 3 and Test 4), adapted from Sedlmeier (1999).

as an inverse function of sample variability. These training attempts were moderately successful, but overall, the training effects were not larger than the differences between the spontaneous solution rates for different types of (frequency-distribution) problems (see Sedlmeier, 1999, pp. 55–60).

Training Programs in High School

The original training programs just described had been used to test different theoretical accounts of people's probabilistic reasoning skills against each other (Sedlmeier, 1999). Each theoretical account was represented by a different type of training program. The success of those program versions described above prompted us to use them as part of a curriculum for German high school students. We wrote a textbook and re-programmed the training programs—which were originally written in Common Lisp and executed on

Macintosh computers (both properties turned out to be obstacles for wider dissemination of the program in Germany)—and added some more programs that cover the whole German high school curriculum on probability theory, including significance tests and confidence intervals based on the binomial distribution (Sedlmeier, 2001; Sedlmeier & Köhlers, 2001). Textbook and software (a Java application) are currently being tested in German high schools (Wassner et al., 2002).

LIMITATIONS OF THE "INTUITIVE APPROACH"

Let me recapitulate. There is evidence that people do badly with some types of probability problems, but when changed a little bit, the problems become much easier. The explanation advanced in this chapter is that the right representational format evokes valid statistical intuitions that help people solve the respective problems. Other factors certainly also play a role but a comparison of the effect sizes found in different attempts at training statistical reasoning (as just mentioned) to those obtained with the current approach lends credence to the hypothesized working of a kind of mostly implicit knowledge I termed *valid intuitions*. I have tried to show that the connection between representational format and the working of valid intuitions is not specific to probabilistic reasoning, but that statistical reasoning is rather a special case. I have also offered an associative learning explanation for the existence of two helpful statistical intuitions, the ratio intuition and the size-confidence intuition. And finally, I have presented evidence for how these intuitions can be used in training statistical reasoning by teaching trainees to translate the information as given in difficult versions of probability problems into versions that can evoke these intuitions.

There is, of course, a limit to the intuitive approach. More complex versions of Bayes' theorem cannot be easily translated into natural frequencies and it is difficult to see how the size-confidence intuition will help in deriving the F distribution. Obviously, the approach advocated here can be very helpful at the beginning stages of learning about probability theory (such as in high school or in the first year of university) but in its current version it does not help much for more complex applications of probability theory. However, first experiences with probability theory may have decisive consequences for a student's later interest and understanding of probability issues.

I now address three other limitations of intuitive probabilistic reasoning. The first concerns the amount of transfer to other tasks obtained in our training studies, the second deals with the context for learning—the schools, and the third is about metalevel considerations: How do we know when solutions to probability problems, intuitive or otherwise, are correct?

Generalization and Transfer

How well did students generalize their solutions to other kinds of problems in our training studies? If we assume that they used valid intuitions, these intuitions were sometimes applied quite narrowly. For instance, in training studies about the probability of conjunctive events and simple conditional probabilities (not discussed here), little generalization was found. If students were trained to solve problems on the probability of conjunctive events they apparently were not able to spontaneously apply the translation skills—translating problem information into a frequency representation similar to the one shown in Figure 16.1—to conditional probability problems and vice versa (Sedlmeier, 2000). However, high school students exposed to Bayesian training as described above were able to generalize the training effect obtained for problems that had the same structure as the training problems to more complex kinds of Bayesian problems (e.g., problems with three instead of two possible outcomes and problems that dealt with extremely low probabilities; see Wassner, 2004). But the question of how well the training generalizes to other kinds of related problems has not been systematically examined yet. Anyway, it would probably be a good idea to heed Lovett and Greenhouse's (2000, p. 4) advice: "give students problems that vary in appearance so their practice will involve applying knowledge and skills in a variety of ways."

Statistical Intuitions In The Classroom

I have cited evidence that even young children possess statistical intuitions. What happens with these intuitions during school? Are they nourished and developed? If one compares the positive conclusions about children's statistical reasoning skills against the negative conclusions in the heuristics and biases literature, one might suspect that school is not that helpful. Indeed, Fischbein (1975) identified school as one of the main reasons for students' diminishing reliance on their intuition of relative frequency. According to him, schools' overwhelming emphasis on deterministic explanations about the world considerably weakens valid statistical intuitions. Recently, Joachim Engel and I went about checking Fischbein's explanation. We gave some probability problems to several classes of fifth-, seventh-, and ninth-graders who had not yet had any formal instruction in probability theory (Engel & Sedlmeier, 2004). One of the problems was this (translated from German):

> A student's final grade in mathematics was "4." Which of the following statements is more likely to apply in his case?
> A. He had a midterm grade of "6."
> B. He has received additional tutoring in the second half of the school year and had a midterm grade of "6."

In this problem the probability of a conjunctive event (tutoring *and* midterm grade of "6") is compared to the probability of a single event (midterm grade of "6").[8] Therefore, Option B cannot be more likely than Option A. Figure 16.8 shows the results for three different kinds of schools.[9] Irrespective of type of school, the solution rates clearly decreased with amount of schooling. Even if one argues that the solution rates for the fifth-graders could be explained by guessing behavior, the systematically increasing percentage of "systematically false guesses" for the seventh- and ninth-graders are in accordance with the assumption that school has a negative effect on students' statistical intuitions. We obtained similar results in a modified version of a task originally used by Piaget and Inhelder (1951/1975). In this task, respondents had to make predictions about the distribution of snowflakes on the tiles of a garage roof by drawing the "snowflakes" on a piece of paper with grids representing the tiles. Again, the number of students who gave deterministic answers (e.g., distributed the snowflakes in a systematic way across tiles) increased with age (Engel & Sedlmeier, 2004; for similar results see Green, 1983, 1991). Informal observations indicated that especially those students who were identified by their math teachers as "excellent" usually failed in the probability tasks. Thus it seems that school might indeed overemphasize a deterministic world view.

What Counts as an Error in Statistical Reasoning?

Finally, when talking about errors and correct solutions in statistical reasoning, some words may be in order about how we determine what an error is. What is the basis for deciding whether statistical reasoning is adequate or not? How does one know that an error has occurred? Usually people's reasoning is compared against some normative rule: "The presence of an error of judgment is demonstrated by comparing people's responses either to an established fact (e.g., that the two lines are equal in length) or to an accepted rule of arithmetic, logic or statistics" (Kahneman & Tversky, 1982b, p. 124). In 1967, Peterson and Beach reviewed research that used probability theory and statistics as a framework for the study of human statistical inference and—contrary to many later researchers—concluded that "the theory of statistical inference can provide a basis for a descriptive theory of imperfect human inference" (p. 20). However, they also cautioned about the pitfalls of blind reliance

[8]In the German school system, grades of 1, 2, 3, 4, and 6 correspond to letter grades of A, B, C, D, and F, respectively.

[9]The school system in the German state of Baden Württemberg, where the study was conducted, divides students up after fourth grade. Some stay at the most basic level (*Hauptschule*), some change to middle level schools (*Realschule*), and others change to high-level schools (*Gymnasium*).

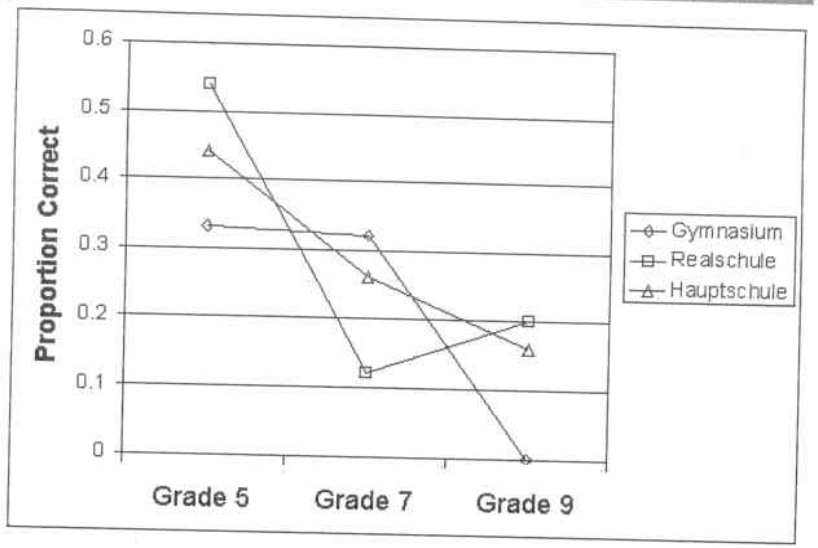


Figure 16.8 Correct solutions in a conjunctive-probability problem over the course of several school years. Results are divided up for different tracks in the German school system (*Hauptschule, Realschule, and Gymnasium*; see Footnote 8).

on theory. They stressed that discrepancies between the statistical model and participants' solutions often arise from the fact that participants' assumptions differ from the presuppositions of the model. So people might deviate from a model assumed by an experimenter but nonetheless reason rationally from the perspective of their own model, which includes additional assumptions. And indeed, the broad conclusion of negative evidence in the judgmental literature has been criticized in several respects along these lines. Apart from the fact that people might have idiosyncratic models (which might be considered normative, given some additional assumptions), statistics does not speak with one voice. Therefore, behavior that appears irrational from the perspective of one statistical theory can be regarded as being perfectly rational from the point of view of another (e.g., Birnbaum, 1983; Cohen, 1981; MacDonald, 1986). Moreover, even when there is no dispute about the right statistical theory, a normative model might be too complex for meaningful use (Sedlmeier & Kilinç, 2004).

Nonetheless, in the vast majority of probability problems that deal with everyday situations, a majority of experts (and people in general) can be

expected to agree on a solution (in some cases, maybe, after some thinking). The problems discussed in this chapter very likely fall into that category. And it is in the treatment of everyday-life problems that people would profit most if valid statistical intuitions are put to use.

REFERENCES

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

Bar-Hillel, M. (1982). Studies of representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 69–98). New York: Cambridge University Press.

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology, 96*, 85–94.

Casscells, W., Schoenberger, A., & Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299*, 999–1000.

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

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences, 4*, 317–331.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*, 187–276.

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.

Dawes, R. M. (1988). *Rational choice in an uncertain world*. San Diego, CA: Harcourt Brace Jovanovich.

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

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). New York: Cambridge University Press.

Engel, J., & Sedlmeier, P. (2004). *Zum Verständnis von Zufall und Variabilität in empirischen Daten bei Schülern* [School students' understanding of chance and variability in empirical data]. *Unterrichtswissenschaft, 32*, 169–191.

Evans, J. St. B. T., & Dusoir, A. E. (1977). Proportionality and sample size as factors in intuitive statistical judgement. *Acta Psychologica, 41*, 129–137.

Evans, J. St. B. T., & Pollard, P. (1985). Intuitive statistical inferences about normally distributed data. *Acta Psychologica, 60*, 57–71.

Fiedler, K., Brinkmann, B., Betsch, R., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base-rate neglect and statistical format. *Journal of Experimental Psychology: General, 129*, 1–20.

Fischbein, E. (1975): *The intuitive sources of probabilistic thinking in children*. Reidel: Dordrecht-Holland.

Fischbein, E. (1994). The interaction between the formal, the algorithmic, and the intuitive components in a mathematical activity. In R. Biehler, R. W. Scholz, R. Sträber, & B. Winkelmann (Eds.), *Didactics of mathematics as a scientific discipline* (pp. 231–245). Dordrecht, the Netherlands: Kluwer.

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., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18*, 253–292.

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

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

Ginossar, Z., & Trope, Y. (1980). The effects of base rates and individuating information on judgments about another person. *Journal of Experimental Social Psychology, 16*, 228–242.

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

Green, D. R. (1983). A survey of probability concepts in 3,000 students aged 11–16 years. In D. R. Grey et al. (Eds.), *Proceedings of the First International Conference on Teaching Statistics* (pp. 766–783). Sheffield, UK: University of Sheffield: Teaching Statistics Trust, University of Sheffield.

Green, D. R. (1991). *A longitudinal study of pupil's probability concepts*. Loughborough, UK: Loughborough University.

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

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

Hogarth, R. (2001). *Educating intuition*. Chicago: University of Chicago Press.

Huber, O. (1993). The development of the probability concept: Some reflections. *Archives de Psychologie, 61*, 187–195.

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. (E. A. Lunzer & D. Papert, Trans.). London: Routledge & Kegan Paul. (Original work published 1959)

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

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. (1982a). The psychology of preferences. *Scientific American, 246*, 160–173.

Kahneman, D. L., & Tversky, A. (1982b). On the study of statistical intuitions. *Cognition, 11*, 123–141.

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

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19*, 1–17.

Klayman, J. & Ha, Y.-W. (1989). Hypothesis testing in rule discovery: strategy, structure and content. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15*, 596–604.

Kuzmak, S. D., & Gelman, R. (1986). Young children's understanding of random phenomena. *Child Development, 57*, 559–566.

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.
Lovett, M. C., & Greenhouse, J. B. (2000). Applying cognitive theory to statistics instruction. *The American Statistician, 54*, 1–11.
MacDonald, R. R. (1986). Credible conceptions and implausible probabilities. *British Journal of Mathematical Psychology, 39*, 15–27.
Mazzoni, G., & Memon, A. (2003). Imagination can create false autobiographical memories. *Psychological Science, 14*, 186–188.
McCormick, J. (1987, August 17). The wisdom of Solomon. *Newsweek*, 24–25.
Murray, J., Iding, M., Farris, H., & Revlin, R. (1987). Sample size salience and statistical inference. *Bulletin of the Psychonomic Society, 25*, 367–369.
Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90*, 339–363.
Norman, D. A. (1988). *The psychology of everyday things*. New York: Basic Books.
Norman, D. A.(1993). *Things that make us smart*. Cambridge, MA: Perseus Books.
Nunes, R., Schliemann, A. D., & Carraher, D. W. (1993). *Street mathematics and school mathematics*. New York: Cambridge University Press.
Olson, C. L. (1976). Some apparent violations of the representativeness heuristic in human judgment. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 599–608.
Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*, 29–46.
Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children* (L. Leake, Jr., P. Burrel, & H. D. Fishbein, Trans.). New York: Norton. (Original work published 1951)
Piattelli-Palmarini, M. (1994). *Inevitable illusions: How mistakes of reason rule our minds*. New York: Wiley.
Reagan, R. T. (1989). Variations on a seminal demonstration of people's insensitivity to sample size. *Organizational Behavior and Human Decision Processes, 43*, 52–57.
Reyna, V. R., & Brainerd, C. J. (1994). The origins of probability judgment: A review of data and theories. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 239–272). Chichester, UK: Wiley.
Ruscio, J. (2003). Comparing Bayes's theorem to frequency-based approaches to teaching Bayesian reasoning. *Teaching of Psychology, 30*, 325–328.
Sedlmeier, P. (1992). *Untersuchungen zu einem Lehr-Lernsystem zum Urteilen unter Unsicherheit* [Studies on a tutorial system concerning judgment under uncertainty]. Unpublished doctoral dissertation, University of Constance, Constance, Germany.
Sedlmeier, P. (1997). BasicBayes: A tutor system for simple Bayesian inference. *Behavior Research Methods, Instruments, & Computers, 29*, 328–336.
Sedlmeier, P. (1998). The distribution matters: Two types of sample-size tasks. *Journal of Behavioral Decision Making, 11*, 281–301.
Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah. NJ: Lawrence Erlbaum Associates.
Sedlmeier, P. (2000). How to improve statistical thinking: Choose the task representation wisely and learn by doing. *Suggestions for the statistics classroom 28*, 227–262.
Sedlmeier, P. (2001). *Statistik ohne Formeln*. In M. Borovcnik, J. Engel, & D. Wickmann (Eds.), *Anregungen zum Stochastikunterricht* (pp. 83–95). Hildesheim: Franzbecker.
Sedlmeier, P. (2002). Associative learning and frequency judgments: The PASS model. In P. Sedlmeier & T. Betsch (Eds.), *Etc.: Frequency processing and cognition* (pp. 137–152). Oxford: Oxford University Press.

Sedlmeier, P. (2005). From associations to intuitive judgment and decision making: Implicitly learning from experience. In T. Betsch & S. Haberstroh (Eds.), *Experience-based decision making* (pp. 89–99). Mahwah, NJ: Lawrence Erlbaum Associates.
Sedlmeier, P. (2006). Intuitive judgments about sample size. In K. Fiedler & P. Juslin (Eds.). *In the beginning there is a sample: Information sampling as a key to understand adaptive cognition* (pp. 53–71). Cambridge: Cambridge University Press.
Sedlmeier, P., & Betsch, T. (2002). (Eds.). *Etc: Frequency processing and cognition*. Oxford: Oxford University Press.
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., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130*, 380–400.
Sedlmeier, P., & Kilinç, B. (2004). The hazards of underspecified models: the case of symmetry in everyday predictions. *Psychological Review, 111*, 770–780.
Sedlmeier, P., & Köhlers, D. (2001). *Wahrscheinlichkeiten im Alltag: Statistik ohne Formeln*. [Probabilities in everyday life: statistics without formula]. Braunschweig: Westermann (textbook with program on CD).
Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: implications for instruction. *Educational Psychology Review, 14*, 47–69.
Swieringa, R., Gibbins, M., Larsson, L., & Sweeney, J. L. (1976). Experiments in the heuristics of human information processing. *Journal of Accounting Research, 4*, 159–187.
Teigen, K. H. (1974). Subjective sampling distributions and the additivity of estimates. *Scandinavian Journal of Psychology, 15*, 50–55.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.
Wassner, C. (2004). *Förderung Bayesianischen Denkens: Kognitionspsychologische Grundlagen und didaktische Analysen*. [Advancing Bayesian reasoning: cognitive psychological foundations and didactical analyses]. Unpublished dissertation, University of Kassel, Germany.
Wassner, C., Martignon, L., & Sedlmeier, P. (2002). Die Bedeutung der Darbietungsform für das alltagsorientiete Lehren von Stochastik [The impact of representational formats on the teaching of statistics for daily life]. *Zeitschrift für Padägogik, 45* (Suppl.), 35–50.
Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes, 47*, 289–312.
Wolfe, C. R. (1995). Information seeking on Bayesian conditional probability problems: A fuzzy-trace theory. *Journal of Behavioral Decision Making, 8*, 85–108.