

Training Conditional and Cumulative Risk Judgements: The Role of Frequencies, Problem-structure and Einstellung

RACHEL McCLOY*, C. PHILIP BEAMAN, BETH MORGAN
and REBECCA SPEED

School of Psychology, University of Reading, UK

SUMMARY

Two experiments implement and evaluate a training scheme for learning to apply frequency formats to probability judgements couched in terms of percentages. Results indicate that both conditional and cumulative probability judgements can be improved in this manner, however the scheme is insufficient to promote any deeper understanding of the problem structure. In both experiments, training on one problem type only (either conditional or cumulative risk judgements) resulted in an inappropriate transfer of a learned method at test. The obstacles facing a frequency-based training programme for teaching appropriate use of probability data are discussed. Copyright © 2006 John Wiley & Sons, Ltd.

People frequently make errors when they come to reason about probability and uncertainty. The negative consequences of these errors have been documented across a range of domains that are of concern for social welfare, most notably in legal and medical contexts. Significant errors arising from fallacious reasoning about probabilities have been demonstrated in, for example, courtroom situations (Faigman, 1999; Koehler, 1992, 1997; Tversky & Kahneman, 1980), HIV counselling (Eddy, 1982; Stine, 1999), medical diagnosis (Casscells, Schoenberger, & Grayboys, 1978), and assessments of risk of reoffending in parole decisions (Gigerenzer, 2002). Additionally, poor probabilistic reasoning may impact upon insurance liabilities, environmental and scientific issues amongst others. An effective means of educating both the general public and relevant experts in reasoning about probability judgement problems would therefore be both useful and desirable.

One class of problems with which people appear to have considerable problems is Bayesian or conditional probability reasoning problems such as:

The probability that a woman aged between 40 and 50 years has breast cancer is 0.8%. If a woman has breast cancer, the probability that this will be successfully detected by a mammogram is 90%. If a woman does not have breast cancer there is a 7% probability

*Correspondence to: R. McCloy, School of Psychology, University of Reading, Earley Gate, Whiteknights, Reading RG6 6AL, UK. E-mail: r.a.mccloy@reading.ac.uk

Contract/grant sponsor: Nuffield Foundation; contract/grant number: H1433900.

of a false positive result on the mammogram. A woman (aged 45) has just tested positive on a mammogram, what is the probability that she actually has breast cancer?

The values in this example conform to the results of a screening mammogram of 26 000 women over 30 years (Gigerenzer, 2002; Kerlikowske, Grady, Barclay, Sickles, & Ernster, 1996a, 1996b). In making judgements of this kind, there is a body of evidence to suggest that people tend to overweight information about hit-rates (here 90%) and underweight information about base-rates (here 0.8%) in making estimates about conditional probabilities (Eddy, 1982; Gigerenzer, 2002). In this case, according to Bayes' theorem, the normatively correct probability that the woman has cancer given a positive test result is only about 9% ($p(\text{cancer}|\text{positive}) = \frac{p(\text{cancer}) \times p(\text{positive}|\text{cancer})}{p(\text{cancer}) \times p(\text{positive}|\text{cancer}) + p(\text{no cancer}) \times p(\text{positive}|\text{no cancer})} = 0.09$). Problems with this *Bayesian* conditional probability structure turn up in a surprising number of situations. These are most usually investigated in a medical context, as in the classic case of breast cancer screening programmes above (see also Eddy, 1982). The structure turns up again in AIDS counselling (Gigerenzer, Hoffrage, & Ebert, 1998) where the problem is concerned with the probability that a client or patient suffers from the condition, given that s/he has tested positive on clinical test and in fact, the problem structure is endemic in all medical contexts concerned with correct diagnosis and representation of the correct probability of some unfortunate result to a patient.

Bayesian conditional probabilities also appear in criminal law; for example, the usefulness of 'associated evidence' (e.g. the general resemblance of the defendant to a description of the offender) can only be assessed by taking into account the base-rate of persons fitting the profile in the general population in order to calculate the conditional probability of a suspect's guilt given a positive 'match' (Koehler, 1992). High-status associated evidence, such as DNA profiles and partial fingerprint matching, is particularly vulnerable to neglect of the base-rate (Koehler, 2001; Koehler, Chia, & Lindsey, 1995). Bluntly, no matter how impressive the underlying forensic test, the probability of guilt must also take into account the base-rates for extra-test sources of possible error. A relatively small base-rate of human error in the laboratory (say 1 in 10 or 10%) if taken into account appropriately could seriously affect judgements of guilt or innocence (Koehler, 1992, 2001; Koehler et al., 1995). Thus, the practical problems—and potential consequences—of erroneous interpretations of probability information can hardly be overstated.

In some situations, however, the neglect of the base-rate focussed upon above is not evident (see Koehler, 1996 for a review) suggesting that the way in which the information is presented is of importance. Recent studies (e.g. Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995) have indicated a means of overcoming fallacies in probabilistic reasoning. Information about uncertainty is normally presented in terms of single event probabilities (e.g. there is a 20% chance of rain today). In studies by Cosmides and Tooby (1996) and by Gigerenzer and Hoffrage (1995) information was presented in what is referred to as 'natural frequency' format, that is, the information is presented as the frequency of occurrence within a defined group or subgroup (e.g. it has rained on 20 out of every 100 previous days like today). The rationale of this is that frequencies occur in the natural environment and are encountered by a process of 'natural sampling' in which both the frequency and the appropriate reference class to which it belongs are immediately evident. Single-event probabilities however, are defined relative to the general population rather than to the appropriate reference class and at the least, require extra calculation. Some

authorities have also questioned the status and validity of ‘one-off’ probability statements (Cosmides & Tooby, 1996; Hacking, 1975, 1990).

Gigerenzer and Hoffrage’s (1995) study showed that performance on Bayesian conditional probability problems like the one described can be significantly improved when the information given is presented in a frequency, rather than a probability format (e.g. 8 of 1000 women have cancer, 7 of these will test positive, 70 of the 992 women without cancer will also test positive, therefore (since we don’t know which group any positive result was drawn from), the chances of someone testing positive actually having cancer must be 7 testing positive out of 7 with cancer plus 70 without cancer, or approximately 9%. This situation is depicted graphically in Figure 1).

From a theoretical perspective, the ‘frequentist’ hypothesis advanced by Gigerenzer and colleagues has been criticized by a number of researchers who have pointed out that

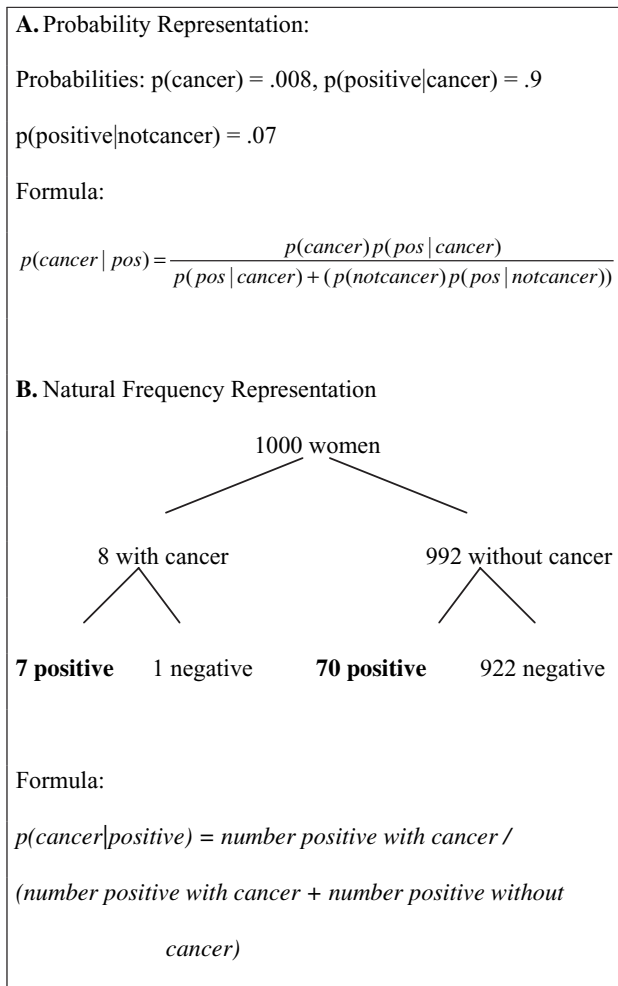


Figure 1. Comparison of the normalized percentage probability and the natural frequency information formats for the breast cancer conditional probability problem. The former must be solved using Bayes’ theorem, the latter may be solved by use of frequency trees

so-called 'natural frequencies' take full advantage of a partitioning of the problem structure that is not available if probability representations are used. Specifically, these researchers note that frequencies *per se* do not hold a monopoly on being expressible relative to an appropriate reference class (e.g. probability can also be expressed in odds or number of chances; Girotto & Gonzalez, 2001) and frequencies may also be expressed in a normalized or non-'natural' form. A number of these researchers (e.g. Girotto & Gonzalez, 2001, Girotto & Gonzalez, 2002; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999; Sloman, Over, Slovak, & Stibel, 2003) have noted that it is possible to explain the results obtained by Gigerenzer and colleagues without assuming a specific module or adaptation for dealing with natural frequencies as do Gigerenzer and Hoffrage (1995). For example, Johnson-Laird et al's (1999) mental models theory of naïve probability judgements is an alternative explanation of such results which does not require the assumption of a specific adaptation. Gigerenzer and others have, in repost, pointed out that the construction and manipulation of arbitrary numbers of mental models can, in itself, be viewed as an activity that requires counting the relative (natural) frequencies of different types of model (Brase, 2002; Hoffrage, Gigerenzer, Krauss, & Martignon, 2002). This issue remains unresolved and may turn ultimately on the vexed question of when a situation may be interpreted as activating the use of natural frequencies *qua* natural frequencies rather than the natural frequency-like properties of (for example) mental models.

From a practical point of view, however, it remains undeniably true that—for whatever reason—situations in which information is presented in natural frequencies rather than probabilities tend to lead to superior performance at conditional probability judgements. One strand of the frequentist research programme has therefore been concerned with demonstrating the superiority of presenting information in this format in a variety of different applied settings (Gigerenzer, 2002; Gigerenzer & Edwards, 2003; Gigerenzer, Hoffrage, & Ebert, 1998; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000). There are two difficulties with this approach, however.

One difficulty is that until such time as all probability information is presented in natural frequencies it remains necessary for people to interpret probability formats. Despite various efforts to inform appropriate professional groups (Gigerenzer & Edwards, 2003; Gigerenzer, Hoffrage, & Ebert, 1998) and the general public (Gigerenzer, 2002) probability or percentage representations of statistical information remain the media of choice in many situations. Slaytor and Ward (1998) reported a wide variety of types of representation of probability information in health-care leaflets concerned with informing women of the risks of breast cancer. More generally, an internet search carried out (on 2nd June, 2005) for the probability value 0.1, using the 'Google' search engine, yielded approximately 1 437 769 hits for frequency information (e.g. 1 in 10) but 198 870 000 hits for probability information in the form of percentages. The frequency search involved 10 terms expressing the value '1 in 10' in both digit and number word forms. The probability search involved two terms expressing the value '10%' in both digit and number word forms—the per cent symbol itself is so common that Google treats it as a function word so the incidence of the term '10%' is not represented by this search.

Amongst professional groups attempting to inform the public, and in the information environment represented by the internet, it is clear that probability and percentage representations are widely used. This difficulty has been partially addressed in a study by Sedlmeier and Gigerenzer (2001; see also Kurzenhauser & Hoffrage, 2002; Ruscio, 2003 and Sedlmeier, 1999) They reported promising results of a teaching regime, based on frequentist research, which aims to tutor students in conditional probability problems by

teaching them to convert from a probability format to a natural frequency representation (Sedlmeier & Gigerenzer, 2001) using a cognitive tutor approach (Anderson, Boyle, Corbett, & Lewis, 1990; Anderson, Corbett, Koedinger, & Pelletier, 1995). We shall return to the details of this teaching regime shortly.

The second problem facing the use of frequentist representations in presenting probability, and particularly risk, information is that although the Bayesian conditional probability problem structure is ubiquitous in medical and criminological contexts it is not the only commonly encountered probability structure. For example, another form of risk commonly encountered in both these contexts is the cumulative probability structure. In one of its simplest forms, a cumulative probability judgement is concerned with the probability that some adverse outcome is avoided, given that the risk on a single occasion is small but the risk is repeatedly taken. Thus, the structure of a *conjunctive* cumulative risk problem (calculating the probability of successfully avoiding an adverse outcome) is as follows:

Suppose that a person living on a floodplain has a 90% probability of not being flooded in any one year. What is the probability that they avoid being flooded at all if they live on the floodplain for 3 years?

The correct solution to conjunctive cumulative risk problems is p^n . In this case $p^n = 0.77$ or 77%. Another obvious example of this kind of risk is the probability of an unwanted pregnancy given the use of a particular contraceptive with less than 100% efficacy. Examples of the complementary *disjunctive* cumulative probability include the probability of experiencing at least one flood having lived on a floodplain for a particular number of years, or the probability of involvement in a car accident (minimal for a single journey, but rather high across a lifetime of such journeys). The correct solution to disjunctive cumulative risk problems is $1 - p^n$. For flood risk, $1 - p^n = 1 - (0.77) = 0.23$ or 23%. People have been shown to make significant and systematic errors on problems of these kinds, both in laboratory studies and in everyday life (Doyle, 1997; Slovic, 2000). For example, in Doyle's (1997) study of cumulative risk, less than 2% of participants (2 out of 128) could correctly judge the level of risk over a period of time. However, a more recent study by McCloy, Byrne, and Johnson-Laird (2006) showed that a natural frequency-like partitioned representation of cumulative risk produced performance superior to that associated with probability representations (which as in the case of Bayesian conditional probability judgements, require knowledge of the relevant statistical formula). The reasons for this are not difficult to see when the situation is represented graphically (see Figure 2) as the same rules for dividing into subsets apply to both the Bayesian judgements where frequency representations have already proven effective and cumulative judgements which were untested prior to McCloy et al's (2006) study.

The first question to be addressed in this study, therefore, is whether the training regime that proved effective for Sedlmeier and Gigerenzer (2001) in training participants to deal with conditional probability judgements relevant to risk-assessments in medical and legal fields is equally effective in training participants to deal with cumulative probability judgements similarly relevant to risk assessments in medicine, healthcare and accident prevention. The second question to be addressed is whether, given training of a particular type, participants are able to generalize across problem structures and employ their newfound expertise (if any) appropriately in a different problem domain. This may be regarded as showing evidence of problem-solving by analogy although this aim is, arguably, quite ambitious. Analogical transfer, even across problems with similar surface

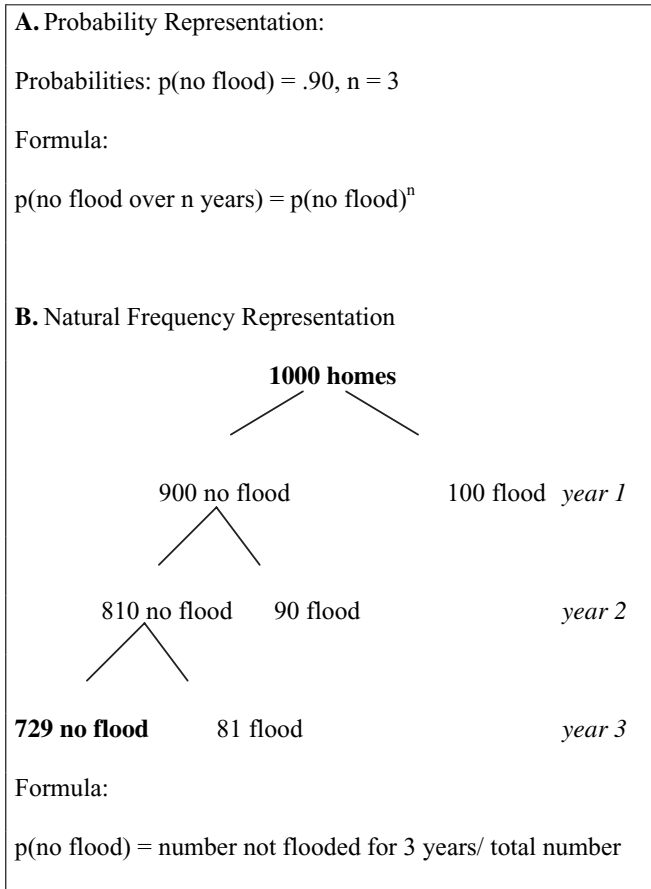


Figure 2. Comparison of the normalized percentage probability and the natural frequency information formats for the flood risk cumulative probability problem

structures, has proven difficult to achieve in the past (Gick & Holyoak, 1983; Holyoak & Koh, 1987; Reeves & Weisberg, 1994). There are three possible outcomes here: either the training will transfer effectively, it will have no effect, or it will be applied inappropriately and actively prevent participants from producing an appropriate response to risk and probability judgements of an unfamiliar type. A third and supplementary question is concerned with the vexed issue of whether representations need to be natural frequencies to be effective or whether being shown the appropriate problem structure is sufficient. Although we do not expect to resolve this third issue within the current study we hope at least to be able to contribute to the debate.

To understand how psychologically-inspired training might improve probability and risk assessments it is necessary to consider the study by Sedlmeier and Gigerenzer (2001). The rationale of this study was that even if the information is not originally presented in a psychologically transparent format, it might nonetheless be possible, with training, for people to transform the information into a more usable representation. This transformation process was taught by demonstrating how a number of different Bayesian conditional probability representations could be converted into frequency representations,

using a tree-structure to clarify the underlying set-subset relationships ('natural samplings') within the data (see Figure 1). Once this transformation process is learned, future probability problems can, in theory, be more easily comprehended regardless of the format in which they are presented as the transformation process can be applied to any conditional probability structure and the answer 'read-off' the bottom of the frequency tree. Sedlmeier and Gigerenzer (2001) found significant and persistent improvements in performance on conditional probability judgements following training on the use of frequency representations for these problems (80–100% correct solution rate) when the pre-training solution rate was only around 10%. Training in the use of Bayes theorem, by contrast, was less immediately effective and solution rate at a later testing-time showed a rapid decline in performance down to near the pre-training, baseline performance level. This frequency teaching regime is therefore extremely promising as an educational tool, but has received, as yet, little testing and has only been applied to training in conditional probability problems of the type outlined above (Kurzenhauser & Hoffrage, 2002; Ruscio, 2003; Sedlmeier, 1999; Sedlmeier & Gigerenzer, 2001).

EXPERIMENT 1

Theoretically, the training system devised by Sedlmeier and Gigerenzer (2001) should be a useful training aid for the kind of everyday risk/benefit problem describable by a cumulative probability structure as these also follow set-subset relations describable in terms of a frequency tree (see also Ruscio, 2003). The current study therefore extends the training system to include cumulative risk judgements in order to determine whether the successes of the training programme generalize beyond conditional probability, and whether the programme is sufficiently powerful to improve the understanding of participants for a variety of different problem types, regardless of the specific problem types they were trained with. As an extra test, we also include training in two formats: frequency and probability. In both cases, the problem structure will be taught in the form of tree diagrams. In this way we will provide a further test of whether problem structure (Johnson-Laird et al., 1999) or problem format (Gigerenzer & Hoffrage, 1995) is the vital element in determining the ease with which people learn how to deal with uncertainty judgements in conditional and cumulative probability problems, and which version (on a practical level) is the most memorable upon re-test. In all cases, because we are concerned with measurable practical benefits of the training, we will be interested only in reasonably large effect sizes comparable to those we expect to see with frequency training of conditional probability problems. Smaller effect sizes may, arguably, be of considerable theoretical significance but of less practical value.

Method

Participants

Sixty-seven undergraduate students from the University of Reading volunteered to take part in the experiment in return for course credit. Ten participants subsequently dropped out at the training stage or failed to return for the retest. This left 57 participants (49 females, 8 males; mean age 20 years, range 18–33 years) in the study—30 in the

frequency training group (15 conditional probability, 15 cumulative probability), and 27 in the probability training group (14 conditional probability, 13 cumulative probability).

Materials and design

All participants were pretested on 12 problems (6 conditional probability and 6 cumulative probability) to provide a baseline before training. These problems were presented in terms of probabilities. All of the problems in this study were presented on a PC running a Windows operating system, using a programme written in Visual Basic. Two groups were trained on conditional probability problems, and two groups were trained on cumulative probability problems. Within each problem type condition, one group received problems framed in terms of probabilities and the other group received problems framed in terms of frequencies. The training problems are described in detail in the next section.

During the training session, all participants received six training problems. The first two of these problems were presented with all of the relevant information completed. Participants received both written and verbal explanations of these problems. Following this, participants were presented with four problems in the same structure, and were required to complete them themselves. During this phase, they received onscreen feedback (see the next section) and could ask for demonstrator help at any time.

Immediately following the training session, all participants were presented with 12 new test problems similar to those received at pre-test (6 conditional probability and 6 cumulative probability). This allowed us to assess both immediate improvement (on trained problems) and immediate transfer (on non-trained problems). This initial testing session (12 pre-test problems, 6 training problems, 12 retest problems) took between 1–1½ hours to complete.

Approximately 1 week after the initial training session, participants returned to the laboratory for a second testing session. In this session, participants again completed 12 problems, 6 of each type (conditional probability, cumulative probability). This allowed us to assess the stability of any improvements (on the trained problems) and the stability of any transfer (on the untrained problems). This retest session took approximately 25 minutes to complete.

Training

All training problems were presented using a tree structure, as in Sedlmeier and Gigerenzer (2001) study 2. The structure of the trees was identical in both the probability and frequency conditions. What differed between these conditions was the information presented at the nodes of the tree. At the top of the tree this showed the size of the relevant reference class in the frequency conditions, and the total probability of events in the probability conditions. At progressively lower levels, the nodes subdivide the reference class, or probability, further, given the information in the problem. Examples of completed trees used in the frequency and probability conditions can be viewed online at http://www.personal.rdg.ac.uk/~sxs98cpb/philip_beaman.htm.

For the first two training problems in each condition, the problems were presented with all of the information filled in. For the remaining four training problems, only the topmost node with the overall reference class or probability was filled in, and participants had to fill in the remaining spaces, and work out the final ratio for themselves on the basis of the information given in the problem. The programme was designed so that participants could not move on from one answer to the next unless the correct answer was entered in the relevant box. If participants entered an incorrect answer and attempted to move on, a

pop-up box informed them that their answer was incorrect, and gave them two options. They could choose either to try again to get the correct answer, or they could choose to have the programme provide the answer for that response box. This applied to all of the answers, both in the tree and in the final ratio. Participants could not move past an answer without making at least one guess.

Results

Analysis was carried out on the 57 participants for which complete baseline, test and re-test data are available. A repeated measures analysis of variance (ANOVA) revealed main effects of the problem type (conditional or cumulative probability), judgements of cumulative risk were more likely to be correct than Bayesian conditional probability judgements, $F(1, 53) = 13.96$, $MSE = 8.46$, effect size (partial η^2) = 0.21, $p = 0.000$. There was a positive effect of training for both conditional and cumulative probability judgements, $F(2, 106) = 65.41$, $MSE = 2.06$, partial $\eta^2 = 0.55$, $p = 0.000$ and an effect of training type (conditional or cumulative), $F(1, 53) = 6.42$, $MSE = 10.52$, partial $\eta^2 = 0.11$, $p = 0.014$ but the main effect of training format (frequency or probability) was not significant, $F < 1$, partial $\eta^2 = 0.001$.

There was no significant interaction between problem type tested and training format, $F < 1$, partial $\eta^2 < .001$, or between time of the testing session and format (frequency or probability), $F < 1$, partial $\eta^2 = 0.01$, suggesting that, collapsed across conditional and cumulative probability problems, there was no substantial difference over time between frequency and probability formats. The interaction between time of testing and training type was marginally significant, however, $F(2, 106) = 3.04$, $MSE = 2.06$, $p = 0.052$, partial $\eta^2 = 0.05$, suggesting that, collapsed across representation formats (probability or frequency), there was a small difference over time between the training types (conditional or cumulative).

Higher-order interactions are most informative as they show how the effects of training type and of training format were affected by the time elapsed since the baseline testing session. These interactions are illustrated in Figures 3 and 4. There were significant interactions between the problem type, the time of test and the *format* of the training, $F(2, 106) = 14.48$, $MSE = 1.71$, $p = 0.000$, partial $\eta^2 = 0.22$. The format (frequency or probability representation) in which the training was presented affected the retention of that training as measured by performance over time with frequency formats providing superior retention for conditional probability problem types. There was also a significant interaction between the problem type, the time of test, and the *type* of training, $F(2, 106) = 61.85$, $MSE = 1.71$, $p = 0.000$, partial $\eta^2 = 0.54$. These two effects seem to be independent of each other but both are dependent upon the type of problem tested (trained or untrained problem type). The higher order interaction between the problem type, time of test, type and format of training was not significant, $F < 1$, partial $\eta^2 = 0.01$.

Further analysis confirms that, for all four groups, performance at 1 week following training was reliably superior to performance on the pre-training baseline test for the trained problem type but there was no evidence of any reliable transfer of understanding across problem types (see Table 1). Paired sample *t*-tests also failed to find any reliable drop in performance between the immediate post-training test session and the 1-week post-training test session for the trained problem types ($p > 0.05$ in all cases before any correction for multiple pairwise comparisons was applied). However, an independent samples *t*-test (1-tailed) revealed a small but statistically significant advantage at 1 week

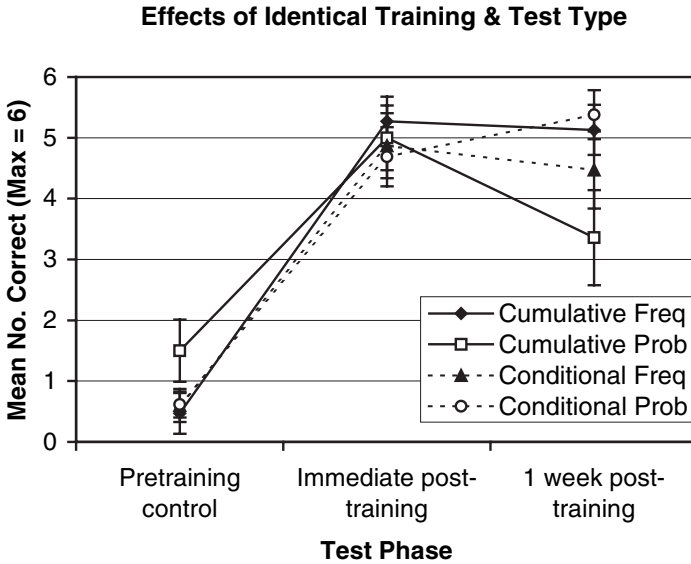


Figure 3. Experiment 1: Effects of tree-format and same-training type on performance at problems of conditional or cumulative probability over time. Bars are standard error

interval for frequency training on conditional probability problems over probability training on the same problem types, $t = 2.01$, $df = 19.82$ (equal variances not assumed), $p = 0.029$ (see Figure 3). As suggested by the main effect of problem type, even before training, participants are significantly more likely to be correct on cumulative probability judgements than on conditional probability judgements ($t = 3.55$, $df = 56$, $p = 0.001$).

Inspection of the rough-work produced by participants as they completed the tasks showed that participants do learn to use tree diagrams of the kind used at training in solving the kinds of probabilistic reasoning problems that we presented. Whilst only 11% of participants showed any evidence of using tree-like diagrams during the pre-test phase of the experiment, 75% used tree diagrams when tested immediately post-training, and 63% used them when retested at a 1 week delay. Participants using tree diagrams post-training used the kinds of tree diagrams on which they had been trained (i.e. frequency trees in the frequency conditions and probability trees in the probability conditions).

As we had not observed any transfer of training across problem types, we additionally examined participants' solutions to the non-trained problems at both immediate retest and at a 1-week delay. We specifically looked to see what proportion of answers in the non-trained problem type could have been produced by an 'Einstellung' effect. Einstellung is a term used to refer to a mental set or a mechanization of thought processes whereby participants 'blindly' follow a familiar procedure even if a more straightforward approach is available (Luchins, 1942; Luchins & Luchins, 1959). In this case, participants who give answers that are consistent with the trained strategy on the untrained problem type for which it was inappropriate would be showing such a mechanization of thought. Table 2 shows the proportion of participants who used the trained strategy on the untrained problems at least once, at both immediate retest and at a 1-week delay.

As Table 2 shows, a small majority of participants use the trained strategy on untrained problems at least once at immediate retest (51%), and this drops only slightly when

Effects of Non-identical Training & Test Type

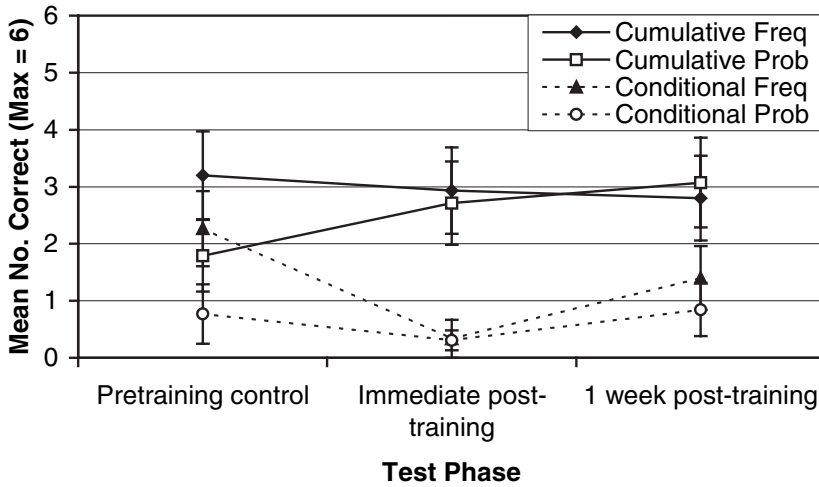


Figure 4. Experiment 1: Effects of tree-format and different-training type on performance at problems of conditional or cumulative probability over time. Bars are standard error

participants are tested at a 1-week delay. Participants appear to be more likely to misapply the trained strategy to the untrained problems in the cumulative conditions than in the conditional conditions, although this difference is only statistically significant at immediate retest (immediate retest: $\chi^2 = 3.96$, $df = 1$, $p = 0.046$; 1 week delay: $\chi^2 = 1.4$, $df = 1$, $p = 0.24$).

So, the experiment shows that, with training, people can improve on both conditional probability and cumulative probability problems. Participants learn to use tree diagrams in order to solve these problems. There is a small retention advantage for training with frequency trees over training with probability trees. We do not find any evidence for transfer of training from one problem type to another. In fact, our results suggest that, as participants display an Einstellung effect—that is, the blind application of the taught strategy to subsequent problems irrespective of their structure. Training on one type of problem may actually hinder solution of another type of problem.

Table 1. Experiment 1: Independent samples *t*- and *p*-values for differences between baseline and 1-week post-training performance on trained and untrained problem types for all training conditions

Training:	Trained problem type		Untrained problem type	
	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>
Conditional, frequency (<i>N</i> = 15)	9.83	0.001	1.87	0.082
Conditional, probability (<i>N</i> = 14)	2.37	0.034	1.61	0.133
Cumulative, frequency (<i>N</i> = 15)	3.18	0.007	1.82	0.09
Cumulative, probability (<i>N</i> = 13)	7.77	0.001	0.56	0.584

Table 2. Experiment 2: Percentage of participants showing an 'Einstellung' effect on untrained problems by condition

	Conditional frequency	Conditional probability	Cumulative frequency	Cumulative probability	Overall
Training condition	<i>n</i> = 15	<i>n</i> = 14	<i>n</i> = 15	<i>n</i> = 13	<i>n</i> = 57
Immediate post-test	40%	36%	67%	62%	51%
1-week delay	40%	36%	53%	54%	46%

Discussion

The results of Experiment 1 replicate those of Sedlmeier and Gigerenzer (2001) and of Ruscio (2003) in showing a positive effect of training using frequency formats and tree structures in aiding conditional probability judgements. The data show very little evidence of a difference between probability tree and frequency tree formats given the appropriate problem structure. However, there is some evidence that conditional training given in a frequency format is more easily retained than the same training given in a probability format. Participants in this experiment received slightly less intensive training than participants in Sedlmeier and Gigerenzer's (2001), which may have contributed towards the lack of transfer. After 1 week's delay participants trained on conditional probability problems using frequency trees showed a mean of 85.5% correct answers compared to Sedlmeier and Gigerenzer's (2001), study 2 median success rate of 93%. Our figure is, however, considerably higher than the 45% success rate immediately following a minimal training session reported by Ruscio (2003).

The results also replicate and extend those of McCloy et al. (2006) in showing that performance on cumulative probability judgements, like performance on conditional probability judgements, can be improved by exposing participants to an appropriate problem structure. Training participants in identifying this problem structure has relatively long-lasting effects—there was little drop-off evident for either problem type over the 1-week interval tested here. The base-line performance was higher in cumulative probability tasks than in conditional probability tasks, but not sufficiently high to affect the conclusions drawn. Superior performance in cumulative probability tasks might, in part, reflect the fact that calculating cumulative probabilities is arithmetically simpler than calculating conditional probabilities, although Doyle (1997) and McCloy et al. (2006) report evidence that participants also reliably choose incorrect strategies in calculating cumulative risk.

In conclusion, the positive result from Experiment 1 shows that with training people can improve on both conditional and cumulative probability judgement problems and that the effects of this training persist, at least over the relatively short interval assessed here. Thus, the use of training to cope with conditional or cumulative probability information that has been presented inappropriately is at least feasible. However, the results of this experiment are not all positive. One negative result of Experiment 1 is that we do not find any evidence for transfer for training from one problem type to another. Although both types of problems can be readily represented and solved using a tree structure, learning a tree structure approach to one type of problem does not facilitate the use of these tree structures for the untrained problem type.

Further examination of participants' responses to the untrained problems shows a second negative result. Our participants display an 'Einstellung' effect on the untrained problems,

regardless of whether their training was conditional or cumulative in nature. That is, they show the blind misapplication of the taught strategy on the untrained problem type. For example, if participants are trained on cumulative probability problems, when presented with conditional probability problems at retest they provide results consistent with a frequency ratio that takes one of the bottom nodes of the frequency tree as numerator and the total sample as denominator (an appropriate strategy for cumulative probability problems). The result represents a serious obstacle to training as a means of dealing with probabilistic information—training on one type of problem may actually hinder solution on another type of problem.

What this suggests overall is that subset knowledge of the kind conveyed in tree structures is easily learned and retained for at least two problem types as demonstrated here. However, it is insufficient for appropriate generalizations to be made when participants are confronted with probabilistic information and given no clue as to how the tree structure is to be traversed. Instead, training encourages more superficial learning of a single pathway that will potentially provide an inappropriate method of making a risk or probability judgement; an *Einstellung* response (see also Mayer & Jackson, 2005, for other ways in which quantitative detail can hinder the development of understanding).

Is there some way in which we can reduce incidences of *Einstellung* responding? If we can encourage participants to consider the structure of the tree diagrams in a less superficial way, perhaps we can not only reduce instances of *Einstellung* responding, but also increase transfer across problem types. If participants consider the structure of the problems that they are asked to solve more deeply they may be more likely to see the differences between the problem types and select the appropriate subsets of the frequency tree required for correct responses. This question is addressed in Experiment 2. To encourage consideration of the structure of the problem, participants were asked multiple questions regarding the tree structures presented to them during the training phase. The rationale for this is that a 'blind' *Einstellung* response merely requires reproduction of a previously learned calculation (Luchins, 1942; Luchins & Luchins, 1959). If participants are required to overtly produce the results of multiple calculations during training and to traverse all possible branches of the tree, this should (a) weaken the influence of a single learned response (e.g. Logan, 1988) and (b) encourage the participants to process the material more deeply and identify the source of the information. We wish to question whether tutoring participants in learning-by-doing within a single type of probability judgement problem is sufficient to allow transfer to a different problem type with the same underlying tree-structure. At the least, it might be expected that further consideration of the underlying problem structure will prevent *Einstellung*-type responses, even if it does not facilitate transfer of training across problem types. Continued *Einstellung*-responses would suggest that a training regime based upon exposing participants to a single problem-type is inappropriate and could actively promote incorrect judgements when presented with untrained problem-types.

EXPERIMENT 2

Method

Participants

Fifty-nine undergraduate students from the University of Reading volunteered to take part in the experiment in return for course credit. The 48 females and 11 males who took part in the

experiment had an average age of 20 years (range 18–37). Thirty participants received training problems followed by only one question (as in Experiment 1; 15 conditional training; 15 cumulative training) and 29 participants received training problems followed by multiple questions (15 conditional training; 14 cumulative training).

Materials and design

In this experiment there were three phases: Pre-test, Training and Immediate Retest. The pre-test and immediate retest phases were the same as in Experiment 1. Participants were presented with 12 problems (6 conditional; 6 cumulative) at pre-test to provide a baseline before training. Following training they were presented with 12 new problems (6 conditional; 6 cumulative), which allowed us to assess both improvement on trained problems and transfer to untrained problems following training. A retest at 1-week's delay was not included in this experiment.

Training

As in Experiment 1, in the training phase participants received six problems. The first two of these problems were again presented with all of the relevant information completed. Participants were then required to complete the remaining four problems themselves. All training problems were presented using a tree structure. In this experiment only frequency trees were employed, as Experiment 1 had revealed no real difference between the two training types (probability; frequency) over the pre-test/immediate retest period. One group of participants was trained on conditional probability problems, the other group were trained on cumulative probability problems.

At the end of each of the training problems, half of the participants were required to calculate a single frequency ratio (___ out of ___) and an associated probability (___%), as in Experiment 1. These are the 'one question' group. For the remainder of the participants (our 'five question' group), training proceeded slightly differently. Following each training problem, the participants were asked five questions about the frequency tree. Each of these questions involved calculating a frequency ratio and an associated probability. The questions were all appropriate to the type of training that participants were receiving. That is, in the conditional probability group they involved the assessment of a conditional probability, and in the cumulative probability group they involved the assessment of a cumulative probability. An example of the 'one question' and 'five questions' versions of the same problem is presented in the Appendix.

Results

A repeated measures analysis of variance (ANOVA) revealed significant main effects of the problem type (conditional or cumulative), judgements of cumulative probabilities were more likely to be correct than those of conditional probabilities, $F(1, 55) = 17.85$, $MSE = 4.41$, effect size (partial η^2) = 0.25, $p = 0.000$,¹ and of the time of the testing session (pre-test or retest), $F(1, 55) = 123.46$, $MSE = 2.31$, partial $\eta^2 = 0.69$, $p = 0.000$, showing a positive effect of training for both conditional and cumulative probability judgements. There was no main effect of training type (conditional or cumulative) or of the number of

¹Again, participants were significantly more likely to be correct on cumulative probability problems than on conditional probability problems even before training ($t = 3.14$, $df = 58$, $p = 0.003$).

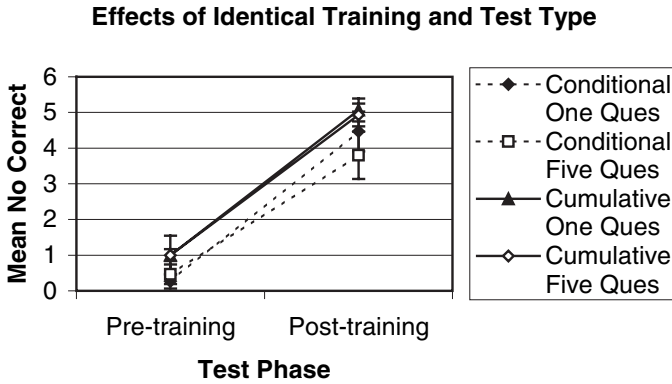


Figure 5. Experiment 2: Effects of same-training type on performance at problems of conditional or cumulative probability. Bars are standard error

questions asked during training (one or five), $F < 1$ in both cases, partial $\eta^2 = 0.03$ and 0.01 , respectively. The main effects can be summarized as follows.

There was no significant interaction between the problem type tested and the number of questions asked, $F < 1$, partial $\eta^2 = 0.02$. There was also no significant interaction between the problem type tested and the time of the testing session, $F(1, 55) = 1.57$, $MSE = 2.31$, partial $\eta^2 = 0.03$, $p = 0.22$, indicating that relative performance on the conditional and cumulative problems did not differ substantially following training. Similarly, there was no significant interaction between the time of the testing session and the number of questions asked, $F < 1$, partial $\eta^2 = 0.02$, suggesting that, collapsed across conditional and cumulative probability problems, there was no substantial difference between the one and five question versions of training over time. There was also no interaction between the number of questions posed during training and the type of training received, $F(1, 55) < 1$.

Again it is the higher order interactions that are most informative. These interactions are illustrated in Figures 5 and 6 below. There was a significant interaction between the type of problem, the time of test, and the type of training received, $F(1, 55) = 72.2$, $MSE = 2.31$, partial $\eta^2 = 0.57$, $p = 0.000$, participants improve on the type of problems on which they

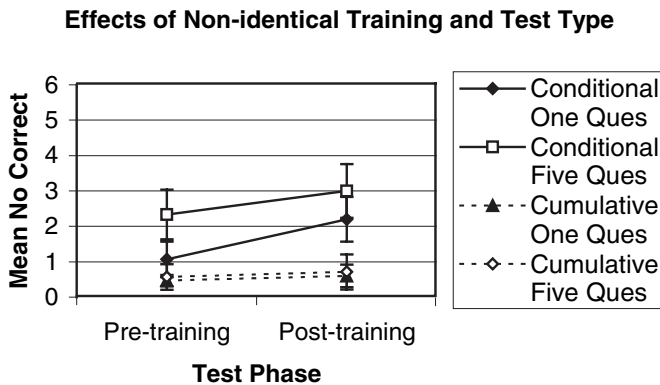


Figure 6. Experiment 1: Effects of different-training type on performance at problems of conditional or cumulative probability. Bars are standard error

have been trained following training. The remaining higher order interactions were not significant, participants did not improve on the untrained problem type, although there is a non-significant increase following conditional training that was not apparent in Experiment 1. This is unaffected by the number of questions posed about each problem during training.

Further analysis confirms that, for the trained problem type in all four training regimes, performance following training was reliably superior to performance on the pre-training baseline test, $t = 7.71$, (conditional training, one question), $t = 5.05$, (conditional training, five questions), $t = 7.2$ (cumulative training, one question) (all $dfs = 14$), $t = 5.56$ (cumulative training, five questions), $df = 13$, all $ps < 0.001$. There was again no evidence for any reliable transfer of understanding across problem types, $t = 2.0$, $df = 14$, $p = 0.07$ for both one and five question conditions (conditional training, cumulative question), $t < 1$, $p > 0.05$ for either one or five question conditions, $df = 14$ and 13 , respectively (cumulative training, conditional question). Forcing participants to traverse all branches of the tree by asking five questions at training rather than one question does not increase transfer from one problem type to another.

Participants' solutions to the untrained problems following training were also examined (see Table 3). As in Experiment 1, we find an 'Einstellung' effect, whereby participants use the trained strategy to answer the untrained problems. In this experiment 63% of participants used the trained strategy on untrained problems at least once. Again, participants seem to be more likely to misapply the trained strategy following cumulative training (72.4%) than following conditional training (53.3%), but this is not significant ($\chi^2 = 3.11$, $df = 1$, $p = 0.08$). Importantly, participants are just as likely to exhibit an Einstellung effect following five-question training (58.6%) as following one-question training (66.7%; $\chi^2 = 2.51$, $df = 1$, $p = 0.11$). Asking more questions during the training phase is not sufficient to prevent the blind misapplication of the learned strategy.

Discussion

The results of Experiment 2 reinforce the conclusions drawn following Experiment 1. There is a confirmed benefit for frequency-tree based training in both conditional and cumulative probability judgements. A non-significant trend for transfer of learning was observed amongst participants who received the conditional training in both one question and five question conditions although this trend was not apparent in Experiment 1 and did not reach statistical significance here. Accompanying this, however, is a statistically reliable but inappropriate transfer of strategy across problem type. Encouraging participants to translate the given percentage-based information formats into more concrete natural frequencies and making the subset relations transparent is demonstrably insufficient to improve performance. The training encourages the 'blind' or rote use of a learned procedure that remains even when attempts are made during training to encourage deeper processing and greater understanding of the steps involved within this procedure. It

Table 3. Experiment 2: Percentage of participants showing an Einstellung response on untrained problems by condition

	Cumulative training	Conditional training
One question	73.3%	60%
Five questions	71.4%	46.7%

was hoped that further questioning during training would lead to a greater awareness of why the procedure works (i.e. the nature of the set-subset relations) rather than, as seems to have been the case, simply reinforcing the procedural aspects of learning that are not readily transferable to different problem types.

GENERAL DISCUSSION

The results of this research programme have produced a number of useful findings. Firstly, the success of a tree-structure based training regime for producing conditional probability judgements (Experiments 1 & 2) has independently replicated the results reported by Sedlmeier and Gigerenzer (2001). The superior stability of frequency representations over time has also been replicated for conditional probability judgements (Experiment 1), however this advantage was not apparent for cumulative probability judgements. Secondly, the results also show that tree-structures are an appropriate means of training participants in making cumulative probability judgements (Experiments 1 & 2), extending the work of McCloy et al. (2006) on the natural frequency-like properties of representations within this problem domain. Thirdly, and perhaps more negatively, the results demonstrate previously unanticipated difficulties in exporting frequency and tree-structure based training schemes into a larger environment in which multiple problem-types exist. As shown in Experiment 2, the Einstellung effect incurred within the training regime is not easily overcome. Such an effect need not be impossible to counteract, of course, although our attempts to do so have not proved successful (Experiment 2) and past research on mechanization of thought strongly suggests that such effects are pervasive and not easily countered (Luchins, 1942; Luchins & Luchins, 1959).

On the theoretical side, the current study has extended work by Sedlmeier and Gigerenzer (2001) and by McCloy et al. (2006) showing the responsiveness of participants to tutoring using formats and structures for which the cognitive system is particularly well adapted (see also Anderson et al., 1990, 1995). These data also add to a currently relatively sparse literature on debiasing and transfer of training in judgement tasks. Importantly, however, training procedures appropriate to a particular problem type do not necessarily improve performance on related but different problem types with the same underlying structure. The results question the extent to which real insight (rather than learned responses) is achieved during this type of training regime and suggest that it may be premature to use such a regime in domains (such as healthcare) where problems of two or more types may occur. The benefits of employing a training regime to improve one type of probability judgement (for example, to estimate healthcare risks of smoking presented earlier) may be outweighed by the costs incurred of forcing an incorrect judgement at another judgement (for example, estimating the conditional risk of breast cancer). Until further research has been carried out to determine when and how appropriate generalizations are made across problem type, it remains appropriate to use tree-structure tutoring systems as training aids for particular problem types (both conditional and cumulative) but it may be inappropriate to use them as a general aid in circumstances where multiple problem types are liable to be encountered. Further research is required to ascertain not only how probability judgements can be made simpler, but also to determine the type of information used to identify an ambiguous problem as being of a particular type (e.g. cumulative or conditional) and react accordingly.

REFERENCES

- Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). Cognitive modelling and intelligent tutoring. *Artificial Intelligence*, *42*, 7–49.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*, 167–207.
- Brase, G. L. (2002). Ecological and evolutionary validity: Comments on Johnson-Laird, Legrenzi, Girotto, Legrenzi and Caverni's (1999) mental-model theory of extensional reasoning. *Psychological Review*, *109*, 722–728.
- Casscells, W., Schoenberger, A., & Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, *299*, 999–1000.
- 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.
- Doyle, J. K. (1997). Judging cumulative risk. *Journal of Applied Social Psychology*, *27*, 500–524.
- Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases*. Cambridge: CUP.
- Faigman, D. L. (1999). *Legal alchemy: The use and misuse of science in the law*. New York: W. H. Freeman.
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38.
- Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster. (Published in the UK as *Reckoning with risk: Learning to live with uncertainty*. London: Penguin).
- Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: From innumeracy to insight. *British Medical Journal*, *27*, 741–744.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *98*, 506–528.
- Gigerenzer, G., Hoffrage, U., & Ebert, A. (1998). AIDS counselling for low-risk clients. *AIDS Care*, *10*, 197–211.
- Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, *78*, 247–276.
- Girotto, V., & Gonzalez, M. (2002). Chances and frequencies in probabilistic reasoning: Rejoinder to Hoffrage, Gigerenzer, Krasuu, and Martignon. *Cognition*, *84*, 252–359.
- Hacking, I. (1975). *The emergence of probability*. Cambridge: CUP.
- Hacking, I. (1990). *The taming of chance*. Cambridge: CUP.
- Hoffrage, U., Gigerenzer, G., Krauss, S., Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, *84*, 343–352.
- Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Medicine: Communicating statistical information. *Science*, *290*, 2261–2262.
- Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory and Cognition*, *15*, 332–340.
- Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Sonino-Legrenzi, M., & Caverni, J-P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, *106*, 62–88.
- Kerlikowske, K., Grady, D., Barclay, J., Sickles, E. A., & Ernster, V. (1996a). Effect of age, breast density, and family history on the sensitivity of first screening mammography. *Journal of the American Medical Association*, *276*, 33–38.
- Kerlikowske, K., Grady, D., Barclay, J., Sickles, E. A., & Ernster, V. (1996b). Likelihood ratios for modern screening mammography: Risk of breast cancer based on age and mammographic interpretation. *Journal of the American Medical Association*, *276*, 39–43.
- Koehler, J. J. (1992). Probabilities in the courtroom: An evaluation of the objections and the policies. In D. K. Kagehiro, & W. S. Laufer (Eds.), *Handbook of Psychology and the Law*. New York: Springer.

- Koehler, J. J. (1996). The base-rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*, 1–17.
- Koehler, J. J. (1997). One in millions, billions, and trillions: Lessons from People vs. Collins (1968) for People vs. Simpson (1995). *Journal of Legal Education*, *47*, 214–223.
- Koehler, J. J. (2001). When are people persuaded by DNA match statistics? *Law and Human Behavior*, *25*, 493–513.
- Koehler, J. J., Chia, A., & Lindsey, S. (1995). The random match probability (RMP) in DNA evidence: Irrelevant and prejudicial? *Jurimetrics Journal*, *35*, 201–219.
- Kurzenhauser, S., & Hoffrage, U. (2002). Teaching Bayesian reasoning: An evaluation of a classroom tutorial for medical students. *Medical Teacher*, *24*, 516–521.
- Logan, G. D. (1988). Towards an instance theory of automatization. *Psychological Review*, *95*, 173–183.
- Luchins, A. S. (1942). Mechanization in problem solving. *Psychological Monographs*, *54*, 248.
- Luchins, A. S., & Luchins, E. H. (1959). *Rigidity of behavior: A variational approach to the effects of Einstellung*. Eugene, OR: University of Oregon Books.
- Mayer, R. E., & Jackson, J. (2005). The case for coherence in scientific explanations: Quantitative details can hurt qualitative understanding. *Journal of Experimental Psychology: Applied*, *11*, 13–18.
- McCloy, R., Byrne, R. M. J., & Johnson-Laird, P. N. (2006). *Understanding cumulative risk*. Manuscript submitted for publication.
- Reeves, L., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. *Psychological Bulletin*, *115*, 381–400.
- Ruscio, J. (2003). Comparing Bayes's theorem to frequency-based approaches to teaching Bayesian reasoning. *Teaching of Psychology*, *30*, 325–326.
- Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, NJ: Erlbaum.
- Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*, 380–400.
- Slaytor, E. K., & Ward, J. E. (1998). How risks of breast cancer and benefits of screening are communicated to women: Analysis of 58 pamphlets. *British Medical Journal*, *317*, 263–264.
- Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, *91*, 296–309.
- Slovic, P. (2000). What does it mean to know a cumulative risk? Adolescents' perceptions of short-term and long-term consequences of smoking. *Journal of Behavioral Decision Making*, *13*, 259–266.
- Stine, G. J. (1999). *AIDS update 1999: An annual overview of acquired immune deficiency syndrome*. Upper Saddle River NJ: Prentice-Hall.
- Tversky, A., & Kahneman, D. (1980). Causal schemata in judgements under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (Vol. 1, pp. 49–72). Hillsdale, NJ: Erlbaum.

APPENDIX

CONDITIONAL

Problem

Ten per cent of the users of the Muscle World gym are female. The gym's records show that if a gym user is female then there is a 70% probability that she uses the gym's tanning facilities. The records also show that if a gym user is male then there is a 20% probability that he uses the gym's tanning facilities. If a gym user is using the tanning facilities, what is the probability that they are female?

One Question

1. If a gym user is using the tanning facilities, what is the probability that they are female?

Five Questions

1. If a gym user is using the tanning facilities, what is the probability that they are female?
2. If a gym user does not use the tanning facilities, what is the probability that they are female?
3. If a gym user is female, what is the probability that they do not use the tanning facilities?
4. If a gym user is male, what is the probability that they do not use the tanning facilities?
5. If a person uses this gym, what is the probability that they use the tanning facilities?

CUMULATIVE

Problem

Primon is a drug used as a last resort in treating a serious illness. With each dose of Primon there is a 25% probability of side effects, that is, there is a 75% probability of avoiding side effects with any one dose. A patient needs to take two doses of Primon. What is the probability that she experiences no side effects at all after taking two doses?

One Question

1. If a patient takes two doses, what is the probability that she experiences no side-effects at all?

Five Questions

1. If a patient takes two doses, what is the probability that she experiences no side-effects at all?
2. If a patient takes two doses, what is the probability that she experiences side-effects at least once?
3. If a patient takes two doses, what is the probability that she experiences side-effects once only?
4. If a patient takes two doses, what is the probability that she experiences the side-effects twice?
5. If a patient takes two doses, what is the probability that she experiences the side-effects once or less?