Problem-Structure and Format in Training Conditional and Cumulative Risk Judgments

Rachel McCloy (r.a.mccloy@reading.ac.uk)
School of Psychology, University of Reading, Reading, U.K., RG6 6AL.

C. Philip Beaman (c.p.beaman@reading.ac.uk)
School of Psychology, University of Reading, Reading, U.K., RG6 6AL.

Abstract

We describe an experiment evaluating a training program designed to help participants interpret probability and risk information presented in a normalized single-event probability format. Both conditional probability and cumulative probability judgments benefited from training in using tree-structures to represent the problem and the set-subset relations involved. The effects of this training persisted over time regardless of whether the training required them to translate the single-event probabilities into “natural” (non-normalized) frequencies or remain with a probability format. However, training did not generalize and was specific to the problem-type tested. A training program for coping with single-event probabilities is therefore feasible for specific problem (or risk judgment) types but may create difficulties when alternative risk or probability judgments are encountered.

Introduction

People frequently make errors when they reason about probability. Negative consequences of such errors have been documented across a range of domains that are of concern for social welfare, most notably in legal and medical contexts (e.g., Gigerenzer, 2002). An effective means of educating both the general public and relevant experts in reasoning about probability judgment problems would therefore be both useful and desirable.

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

The probability that a woman aged between 40 and 50 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 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?

In making judgments 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%. 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.

Gigerenzer and Hoffrage’s (1995) study showed that performance on Bayesian conditional probability problems like the one described above 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 out of 7 plus 70, or approximately 9%.

From a theoretical perspective, the “frequentist” hypothesis advanced by Gigerenzer and colleagues has been criticized by a number of researchers who point out that “natural frequencies” take advantage of a partitioning of the problem structure. 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).

From a practical point of view, however, it remains undeniably true that situations in which information is presented in natural frequencies rather than probabilities tend to lead to superior performance at conditional probability
judgments. One strand of the frequentist research program 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; 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 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. 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. For example, an internet search (on 7/16/04) for the probability value .1, using the Google search engine, yielded approximately 445,040 hits for frequency information (e.g., 1 in 10) but 904,941,000 hits for probability information in the form of percentages. This difficulty has been partially addressed in a study by Sedlmeier and Gigerenzer (2001) 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.

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 is the cumulative probability structure. In its simplest form, a cumulative probability judgment 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 the 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?1

Obvious examples of this kind of risk include the probability of involvement in a car accident (minimal for a single journey, but rather high across a lifetime of such journeys) and the probability of an unwanted pregnancy given the use of a particular contraceptive with less than 100% efficacy. People have been shown to make significant and systematic errors on problems of this kind, both in laboratory studies and in everyday life (Doyle, 1997; Slovic, 2000). In Doyle’s (1997) study of cumulative risk, less than 2% of participants (2 out of 128) could correctly judge the level risk over a period of time. However, a more recent study by McCloy, Byrne and Johnson-Laird (2005) showed that a natural frequency-like representation of cumulative risk produced performance superior to that associated with probability representations (which requires knowledge of the relevant statistical formula). The reasons for this are not difficult to see as the same rules for dividing into subsets apply to both the Bayesian judgments where frequency representations have already proven effective and cumulative judgments which were untested prior to McCloy et al’s (2005) 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 judgments is equally effective in training participants to deal with cumulative probability. The second question to be addressed is whether, given training, participants are able to generalize across problem structures and employ their newfound expertise appropriately to a different problem type. 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 probability judgments 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 an appropriate problem structure is sufficient.

The rationale of the Sedlmeier and Gigerenzer (2001) study was that even if the information is not originally presented in a psychologically transparent format, it might nonetheless be possible, with training, 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. Once this transformation process is learned, future probability problems can, in theory, be more easily comprehended regardless of the format in which they are encountered. The transformation process can be applied to any conditional probability structure and the answer “read-off” the bottom of the frequency tree (see Appendix, Figure A1 for an example frequency tree). Sedlmeier and Gigerenzer (2001) found significant and persistent improvements in performance on conditional probability judgments following training on the use of frequency representations for these problems. 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

---

1 The correct solution to such problems is $p^n$. In this case $p^n = 0.77$ or 77%.
and has only been applied to training in conditional probability problems of the type outlined above.

Experiment

Theoretically, the training system devised by Sedlemeier 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 Appendix Figure A2). The current study therefore extends the training system to include cumulative risk judgments in order to determine whether the successes of the training program generalize beyond conditional probability, and whether the program 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 trees and probability trees. In this way we will provide a further test of whether problem structure (e.g., Girotto & Gonzalez, 2001) or problem format (Gigerenzer & Hoffrage, 1995) is the vital element in determining the ease with which people learn how to deal with uncertainty judgments in conditional and cumulative probability problems, and which version is the most memorable upon re-test.

Participants

67 undergraduate students from the University of Reading volunteered to take part in the experiment in return for course credit. 10 participants subsequently dropped out at the training stage or failed to return for the retest. This left 57 participants (49 females, 8 males) 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 program 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.

During the training session, all participants received six training problems in a tree structure. 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 both intermediate calculations and final answers themselves. During this phase, they received onscreen feedback and could ask for demonstrator help at any time.

Immediately after the training session, all participants were presented with 12 new test problems similar to those received at pretest (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).

Approximately one 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).

Training

All training problems were presented using a tree structure, as in Sedlemeier 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. The top of the tree 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 (see Appendix Figure A3).

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 program was designed so that participants could not move on from one answer to the next unless the correct answer was given. 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, or they could choose to be provided with the answer. This applied to all of the missing information: intermediate calculations represented as nodes on the tree and the final ratio. Participants started filling in information from the top of the tree and could not move past nodes without making at least one attempt to answer. For all training and testing sessions paper and pencil and electronic calculators were made available for participants.

Results

A repeated measures analysis of variance (ANOVA) revealed main effects of the problem type (conditional or cumulative probability), $F(1, 53) = 13.96, p < .001$, the time...
of the testing session, $F(2, 106) = 65.41, p < .001$ and of training type (conditional or cumulative), $F(1, 53) = 6.42, p < .02$ but the main effect of training format (frequency or probability) was not significant, $F < 1$. The main effects can be summarized as follows: judgments of cumulative risk were more likely to be correct than Bayesian conditional probability judgments. There was a positive effect of training for both conditional and cumulative probability judgments but no effect of the format (frequency or probability) in which the training was given.

All but one of the two-way interactions were non-significant, however, the interaction between time of testing and training type was marginally significant, $F(2, 106) = 3.04, p = .052$, 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 1 and 2. There were significant interactions between the problem type, the time of test and the format of the training, $F(2, 106) = 14.48, p < .001$, and between the problem type, the time of test, and the type of training, $F(2, 106) = 61.85, p < .001$. The higher order interaction between the problem type, time of test, type and format of training was not significant, $F < 1$. These results can be summarized as follows: the format (frequency or probability representation) in which the training was presented affected the retention of that training as measured by performance over time, as did the type of training (conditional or cumulative). These two effects seem to be independent of each other but both are dependent upon the type of problem tested (conditional or cumulative probability judgment).

Further analysis confirms that, for all four groups, performance at 1 week 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 one-week post-training test session for the trained problem types ($p > .05$ in all cases). However, an independent samples $t$-test (1-tailed) revealed a small but statistically significant advantage at 1 week interval for frequency training on conditional probability problems over probability training on the same problem types, $t = 2.01, df = 19.82$ (unequal variances), $p < .03$.

Table 1: Independent samples $t$- and $p$ values for differences between baseline and 1 week post-training performance on trained and untrained problem types

<table>
<thead>
<tr>
<th>Training Type</th>
<th>Trained Problem Type</th>
<th>Untrained Problem Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond. Freq. (N=15)</td>
<td>9.83 .001**</td>
<td>1.87 .082</td>
</tr>
<tr>
<td>Cond. Prob. (N=14)</td>
<td>2.37 .034*</td>
<td>1.61 .133</td>
</tr>
<tr>
<td>Cum. Freq. (N=15)</td>
<td>3.18 .007**</td>
<td>1.82 .09</td>
</tr>
<tr>
<td>Cum. Prob. (N=13)</td>
<td>7.77 .001**</td>
<td>.56 .584</td>
</tr>
</tbody>
</table>

(* values are significant at the .05 level and ** values are significant at the .01 level).

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 one-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 (e.g., Luchins & Luchins, 1959), where participants “blindly” follow a familiar procedure. In this case, that would be participants who give answers that are consistent with the trained strategy on the untrained problem type, for which it was inappropriate.
For example giving a conditional probability result to a cumulative probability question and vice versa.

As Table 2 shows, over 50% of participants use the trained strategy on the untrained problem type at least once at immediate retest, and this drops only slightly when participants are tested at a one-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 not statistically significant ($\chi^2 > 0.05$).

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Immediate Post-test</td>
<td>40% 36%</td>
<td>67% 62%</td>
<td>51%</td>
<td></td>
<td>n = 15</td>
</tr>
<tr>
<td>One-week Delay</td>
<td>40% 36%</td>
<td>53% 54%</td>
<td>46%</td>
<td></td>
<td>n = 13</td>
</tr>
</tbody>
</table>

**General Discussion**

These data replicate those of Sedlmeier and Gigerenzer (2001) in showing a positive effect of training using frequency formats and tree structures in aiding conditional probability judgments. The data also show very little difference between probability and frequency formats provided the structure is appropriate, although there is some evidence that conditional training given a frequency format is better retained than the same training given a probability format. More importantly, the results also replicate and extend those of McCloy et al. (2005) in showing that performance on cumulative probability judgments, like conditional probability judgments, can be improved by exposing participants to an appropriate problem-structure. Training participants in identifying this problem-structure has demonstrated here. However it is insufficient for appropriate generalizations to be made when participants are confronted with probabilistic information and given no cue concerning 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 judgment, an Einstellung response. Further experiments are required to determine how serious this potential problem is and what methods might exist to overcome it in a larger-scale training program.

**Acknowledgments**

This research was supported by grant SGS/00897/G from the Nuffield Foundation. Thanks to Andy Fitzsimmons and Natalie Lynch for their help in programming and running the experiment.

**References**


Appendix: Example Tree Structures

1000 People

8 disease

992 no disease

7 positive

1 negative

922 positive

70 negative

Figure A1: Conditional probability of suffering from the disease given a positive test result = $\frac{7}{7 + 70} = 9\%$.

100 Women

10 pregnancies

90 no pregnancy

1 pregnancy

9 no pregnancies

81 no pregnancy

Figure A2: Cumulative risk of at least one unwanted pregnancy over 2 years if contraception is 90% efficient = $(10 + 9) / 100 = 19\%$.

Figure A3: Screendump showing example of partially completed probability tree for conditional probability condition. The problem here is, on completion of the tree, to calculate the probability that a person buying a stereo owns a top range car (Answer: 0.045 out of (0.045 + 0.076) = 37.2%).