Facilitating Normative Judgments of Conditional Probability: Frequency or Nested Sets?

Kimihiko Yamagishi

Graduate School of Decision Science & Technology, Tokyo Institute of Technology, Japan

Abstract. Recent probability judgment research contrasts two opposing views. Some theorists have emphasized the role of frequency representations in facilitating probabilistic correctness; opponents have noted that visualizing the probabilistic structure of the task sufficiently facilitates normative reasoning. In the current experiment, the following conditional probability task, an isomorph of the "Problem of Three Prisoners" was tested. "A factory manufactures artificial gemstones. Each gemstone has a $\frac{1}{3}$ chance of being blurred, a $\frac{1}{3}$ chance of being cracked, and a $\frac{1}{3}$ chance of being clear. An inspection machine removes all cracked gemstones, and retains all clear gemstones. However, the machine removes $\frac{1}{2}$ of the blurred gemstones. What is the chance that a gemstone is blurred after the inspection?" A 2×2 design was administered. The first variable was the use of frequency instruction. The second manipulation was the use of a roulette-wheel diagram that illustrated a "nested-sets" relationship between the prior and the posterior probabilities. Results from two experiments showed that frequency alone had modest effects, while the nested-sets instruction achieved a superior facilitation of normative reasoning. The third experiment compared the roulette-wheel diagram to tree diagrams that also showed the nested-sets relationship. The roulette-wheel diagram outperformed the tree diagrams in facilitation of probabilistic reasoning. Implications for understanding the nature of intuitive probability judgments are discussed.

Key words: probability judgment, nested sets, frequency judgment

Recent research on probabilistic judgment shows opposing arguments regarding the effectiveness of presenting uncertainty information in frequency representations. Gigerenzer (1991, 1994) may be regarded as a champion of one extreme side of this contrast, asserting that so-called *cognitive illusions* "disappear" when statistical tasks are presented in frequency representations. In disagreement, others have remarked that logical contradictions are intrinsic to frequency judgments (Denes-Raj & Epstein, 1994; Griffin & Buehler, 1999; Macchi, 2000; Yamagishi, 1994a, 1994b, 1997a, 1997b), casting doubts as to whether frequency judgments necessarily adhere to applicable statistical norms.

In this study, the focus is on the failure to incorporate base rates in conditional probability judgments (e.g., Kahneman & Tversky, 1973) among the judgmental phenomena extensively discussed by Gigerenzer (1991, 1994). Particularly, an isomorph of the "Problem of Three Prisoners" was used. In its original form (Mosteller, 1965), the problem reads as follows:

Three men, A, B, and C, were in jail. A knew that one of them was to be set free and the other two were to be executed. But he didn't know who was the one to be spared. To the jailer, who did know, A said, "Since two out of the three will be executed, it is certain that either B or C will be, at least. You will give me no information about my own chances if you give me the name of one man, B or C, who is going to be executed." Accepting this argument after some thinking, the jailer said, "B will be executed." Thereupon, A felt happier because now either he or C would go free, so his chance had increased from $\frac{1}{3}$ to $\frac{1}{2}$. This prisoner's happiness may or may not be reasonable. What do you think?

Preliminary reports of this work have been presented at the 2001 meetings of Subjective Probability, Utility, and Decision Making Conference, the 42nd annual meeting of the Psychonomic Society, and the 65th annual conference of Japanese Psychological Association. This work was supported by Grant No. 14510093 from the Japanese Ministry of Science, Education, and Culture. I would like to thank Ulrich Hoffrage, Shin'ichi Ichikawa, Yuji Itoh, and Steven Sloman for helpful suggestions.

DOI: 10.1027//1618-3169.50.2.97

Normatively, this task should be approached as a problem of conditional probability. Let $p(A)$, $p(B)$, and $p(C)$ denote the chances that A, B, and C, respectively, will survive. It would be reasonable to assume that $p(A) = p(B) = p(C) = 1/3$. Furthermore, assume that the jailer always tells the truth and he has no preference between answering B and C if both are to be executed. Let d denote the datum, namely the jailer saying that "B will be executed." Let $p(d|A)$ denote the probability of obtaining the datum given that A was to be freed, and analogously for $p(d|B)$ and $p(d|C)$. Then, $p(d|A) = 1/2$, $p(d|B) = 0$, and $p(d|C) = 1$ would follow. Thus, the chance of A's survival can be calculated by Bayes' theorem in the following formula:

$$p(A|d) = \frac{p(d|A)p(A)}{p(d|A)p(A) + p(d|B)p(B) +, p(d|C)p(C)}. \quad (1)$$

Therefore, $p(A|d)$ is derived as:

$$p(A|d) = \frac{(1/2)(1/3)}{(1/2)(1/3) + 0 \cdot (1/3) + 1 \cdot (1/3)} = \frac{1}{3}. \quad (2)$$

Previous research has documented the difficulty of this problem. In the experiments of Shimojo and Ichikawa (1989), the normative answer rates were as low as 0–18%. They pointed out that the difficulty lies in incorporating the prior probability into judgment, namely the numerator in Formulae (1) and (2), and instead people rely on other heuristics to replace the calculation of $p(d|A)p(A)$. Gigerenzer (1994, p. 164), however, argued that this sort of conditional probability task could be "easily digested" when translated into frequency representations because frequency makes the relationship between the prior and the posterior probabilities immediately understandable. As a frequency version of the Problem of Three Prisoners, Itoh (1991) proposed the following "Gemstone Problem":

A factory manufactures 1200 artificial gemstones daily. Among the 1200, 400 gemstones are blurred, 400 are cracked, and 400 contain neither. An inspection machine removes all cracked gemstones, and retains all clear gemstones. However, the machine removes half of the blurred gemstones. How many gemstones pass the inspection, and how many among them are blurred?

The Gemstone Problem's isomorphism to the Problem of Three Prisoners should be clear if the following correspondence is agreed upon: $p(A)$ to p(blurred), $p(B)$ to p(cracked), and $p(C)$ to p(clear); and $p(A|d)$ to p(blurred|inspection). Moreover, in the frequency description, it could be easily calculated that 400 clear gemstones and 200 blurred ones remain after the inspection. Therefore,

$$p(\text{blurred}|\text{inspection}) = \frac{200}{200 + 400} = \frac{1}{3}. \quad (3)$$

Two Competing Accounts: Frequency Versus Nested Sets

One essential question raised is: Why does the Problem of Three Prisoners seem very difficult, while the Gemstone Problem in the frequency instruction seems substantially easier to work on? In contrast to Gigerenzer's argument that frequency makes it easier to incorporate the prior probability into judgment, Mellers and McGraw (1999) provided another explanation. According to Mellers and McGraw, frequencies allow people to visualize "nested sets," namely, subsets relative to larger sets, in the task structure. In the Gemstone Problem, "nested sets" refers to the proportion of blurred and inspected gemstones in the total number of blurred gemstones. Once the nested sets, that among 400 blurred gemstones, 200 pass the inspection, become clear, then one can readily answer the Gemstone Problem or carry out the conditional probability calculation as in Formula (3). An important difference of the nested-sets argument from Gigerenzer's argument is that the logic of nested sets does not necessarily require frequency in grasping the relationship between p(blurred) and p(inspected and blurred).

In this study, the Gemstone Problem was used to contrast the effectiveness of the frequency instruction to a nested-sets instruction. As a nested-sets instruction that does not necessarily call for attention to frequency, a roulette-wheel diagram was adopted (Ichikawa, 1989) that shows the relationship between the prior and the posterior probabilities (see Figure 1). In Figure 1, the inner circle labeled "Production" shows the breakdown of the prior probabilities for the blurred, cracked, and clear gemstones. The rim labeled "Inspection" shows the distributions of probabilities that a gemstone is clear or blurred after the inspection. The set-inclusion relationship between the prior and the posterior probabilities shows up in an obvious fashion in Figure 1. Ichikawa (1989) suggested that such diagrammatic representations would be effective in Bayesian probabilistic reasoning (see

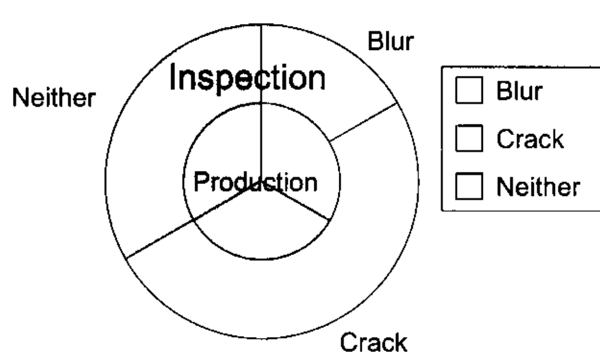

Figure 1. Roulette-wheel diagram used in Experiment 1.

Sedlmeier (1999) for a more recent attempt to use visual aids in Bayesian inference tasks). Although Ichikawa did not mention the idea of nested sets, the effectiveness of the diagrammatic representation obviously lies in graphically showing the nested sets of p(blurred) and p(inspected and blurred). The following experiments were administered in a 2×2 design created by having frequency instruction and nested sets instruction as independent variables. In using the roulette-wheel diagram, the author's purpose was to examine if frequency instruction facilitates normative Bayesian reasoning as effectively as nested-sets instruction.

Experiment 1

The Gemstone Problem

The following was presented as a probability version of the Gemstone Problem:

A factory manufactures artificial gemstones. Each gemstone has a ⅓ chance that it is blurred, a ⅓ chance that it is cracked, and a ⅓ chance that it contains neither. An inspection machine removes all cracked gemstones, and retains all clear gemstones. However, the machine removes ½ of blurred gemstones. What is the chance that a gemstone is blurred after the inspection?

The aforementioned frequency version of the Gemstone Problem was used in the frequency instruction. Participants in the nested sets conditions were provided with the diagram in Figure 1 along with the verbal description of the Gemstone Problem (the actual diagrams used in the experiments used Japanese words instead of English in the figures). They were instructed that: The inner circle labeled "Production" shows the proportion of the blurred, cracked, and clear gemstones; and the rim labeled "Inspection" shows the proportion of clear or blurred gemstones, after the inspection.

Independent Variables

The Gemstone Problem was presented either in the frequency or in the probability version. In addition, the task was presented either with or without the diagram shown in Figure 1. Thus, $2 \times 2 = 4$ conditions were set up.

Dependent Variable

Participants who received the probability version were instructed to answer with the intuitive probability. Responses from the probability version conditions were coded correct if the answer indicated "⅓" or equivalent responses (such as "0.333 …"). Participants in the frequency-version conditions answered with the number of gemstones that passed the inspection and the number of gemstones that are blurred after the inspection. Responses from the frequency-version conditions were coded correct if the response indicated the former and the latter numbers in question as 600 and 200, respectively.

Participants

One hundred and sixty-eight Japanese undergraduates participated. They were enrolled in an introductory psychology course and participated as a partial fulfillment of the course credit. Each participant was randomly assigned to one of the four conditions.

Procedure

Data were gathered in a group setting in one classroom. Each participant was provided with a questionnaire booklet and was instructed to work on the booklet individually. The booklet contained the Gemstone Problem in one of the four conditions, along with other filler tasks. The participants received all oral and written instructions in Japanese.

Results and Discussion 1

Two-tailed tests were used throughout data analyses in this paper.

Table 1 shows the percentages of correct responses in each condition. The proportion of correct responses between Group 3 and Group 4 showed a significant difference ($z = 2.66$, $p < .01$), demonstrating the effectiveness of the frequency instruction. All statistical tests of proportional differences between groups in this paper were performed by using Goodman's (1964) test statistic,

$$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}}. \quad (4)$$

Table 1. The Percentages of Correct Responses in Experiment 1

	Frequency Instruction	Probability Instruction
Diagram Present	Group 1 ($n_1 = 47$) 72.3%	Group 2 ($n_2 = 43$) 69.8%
Diagram Absent	Group 3 ($n_3 = 40$) 42.5%	Group 4 ($n_4 = 38$) 18.4%

However, Group 1 vs. Group 2 did not show a reliable difference ($z = .261$). The correct response rates in Group 1 and Group 2 exceeded the correct response rate in Group 3 ($z = 3.03$ and 2.98, respectively, both $p < .01$). Thus, the frequency instruction was only somewhat successful, and the roulette-wheel diagram alone more effectively facilitated proper probabilistic reasoning than the frequency instruction.

A proponent of frequency instruction may suggest that the insignificance between Group 1 and Group 2 is caused by a ceiling effect. Yet, data are inconsistent with this rebuttal. If a ceiling effect were in effect, the correct response rates for Group 1 and Group 2 would encompass 100% within their margins of error. In Groups 1 and 2, the 95% confidence intervals for the correct response proportion are 72.3 ± 12.8% and 69.8 ± 11.9%, respectively, with neither reaching 100%.

However, a possible criticism is that, in Experiment 1, a problem solver can draw a seemingly normative solution simply by substituting the prior probability (⅓) to the required answer (⅓) (Shimojo & Ichikawa, 1989). In response to this possible criticism, Experiment 2 varied the distribution of the prior probabilities so that the posterior probability does not correspond with the prior probability.

Experiment 2

In the probability instruction of Experiment 2, the prior probabilities for the blurred, cracked, and clear gemstones were described as ¼, ¼, and ½, respectively. Therefore, the probability that a gemstone after the inspection is blurred is given as:

$$p(\text{blurred}|\text{inspection}) = \frac{(\tfrac{1}{2}) \cdot p(\text{blurred})}{(\tfrac{1}{2}) \cdot p(\text{blurred}) + 0 \cdot p(\text{cracked}) + 1 \cdot p(\text{clear})}$$

$$= \frac{(\tfrac{1}{2})(\tfrac{1}{4})}{(\tfrac{1}{2})(\tfrac{1}{4}) + 0 \cdot (\tfrac{1}{4}) + 1 \cdot \tfrac{1}{2}} = \frac{1}{5}. \quad (5)$$

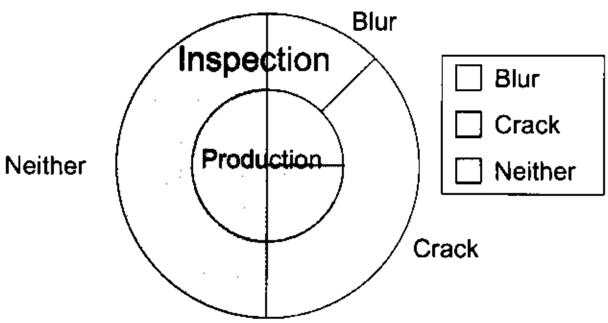

Figure 2. Roulette-wheel diagram used in Experiment 2.

Note that, unlike Experiment 1, the conditional probability does not equal any of the prior probabilities. Figure 2 illustrates how this normative solution may be obtained by showing the nested sets. Figure 2 was used in the nested-sets instruction conditions in Experiment 2.

The Gemstone Problem

The following was presented in the frequency instruction conditions:

A factory manufactures 1200 artificial gemstones daily. Among the 1200, 300 gemstones are blurred, 300 are cracked, and 600 contain neither. An inspection machine removes all cracked gemstones, and retains all clear gemstones. However, the machine removes half of blurred gemstones. How many gemstones pass the inspection, and how many among them are blurred?

In the probability instruction conditions, the prior probabilities for the blurred, cracked, and clear gemstones were described as ¼, ¼, and ½, respectively, rather than as frequencies. Participants in the probability instruction conditions were asked to answer with the conditional probability that a gemstone is blurred after the inspection. Participants in the nested-sets conditions were provided with the diagram in Figure 2 along with the verbal description of the Gemstone Problem. They received instructions that were analogous to the nested-sets conditions in Experiment 1 regarding the inner circle and the rim in Figure 2.

Independent Variables

The Gemstone Problem was presented either in the frequency instruction or in the probability instruction. In addition, the task was presented either with or without the diagram shown in Figure 2. Thus, $2 \times 2 = 4$ conditions were set up.

Dependent Variable

Participants who received the probability instruction were asked to answer with the intuitive probability. Responses from the probability version conditions were coded correct if the answer indicated "⅕" or equivalent responses (such as "0.2"). Participants in the frequency-version conditions answered with the number of gemstones that passed the inspection and the number of gemstones that are blurred after the inspection. Responses from the frequency-version conditions were coded correct if the response indicated the former and the latter numbers in question as 750 and 150, respectively.

Participants and Procedure

One hundred and seventy-one Japanese undergraduates participated. They were enrolled in an introductory psychology course and participated as a partial fulfillment of the course credit. The rest of the procedure was the same as in Experiment 1.

Results and Discussion 2

Table 2 shows the percentages of correct responses in each condition. The proportion of correct responses between Group 7 and Group 8 showed a significant difference ($z = 4.82$, $p < .001$), demonstrating the effectiveness of the frequency instruction. However, Group 5 vs. Group 6 did not show a reliable difference ($z = -.47$). The correct response rates in Group 5 and Group 6 surpassed the correct response rate in Group 7 ($z = 2.15$ and 2.64, respectively, both $p < .05$). Hence, although the frequency instruction was modestly successful, the roulette-wheel diagram alone more effectively facilitated normative reasoning than did the frequency instruction.

Table 2. The Percentages of Correct Responses in Experiment 2

	Frequency Instruction	Probability Instruction
Diagram Present	Group 5 ($n_5 = 43$) 76.7%	Group 6 ($n_6 = 42$) 80.9%
Diagram Absent	Group 7 ($n_7 = 45$) 55.6%	Group 8 ($n_8 = 41$) 12.2%

As in Experiment 1, theorists in defense of the frequentist viewpoint may rebut that the insignificant difference between Group 5 and Group 6 is the result of a ceiling effect. Yet, the data are inconsistent with this rebuttal. In Group 5 and Group 6, the 95% confidence intervals for the correct response proportion are 76.7 ± 12.6% and 80.9 ± 11.9%, both without reaching 100%.

Experiment 3

Regarding the results from Experiments 1 and 2, the following may arise as a possible criticism: The nested sets groups (Groups 1, 2, 5, and 6) always had the advantage of having the roulette-wheel diagram as a visual aid. Therefore, the results that seemingly demonstrate the advantage of the nested-sets hypothesis over the frequentist hypothesis may really show the difference between solving the problem with or without a proper visual aid. Experiment 3 was designed to respond to this concern by adding a suitable visual representation to the problems administered without the roulette-wheel diagram in Experiments 1 and 2.

Figure 3 shows the frequency tree used in Experiment 3. This tree closely resembles the frequency tree used by Sedlmeier and Gigerenzer (2001) in their Bayesian reasoning task to judge the conditional probability of having breast cancer, given a positive mammogram. The frequency tree proved effective in facilitating the appropriate Bayesian reasoning. More than 90% of Sedlmeier and Gigerenzer's participants solved the mammography problem correctly (Sedlmeier & Gigerenzer, 2001, Study 2). It should be noted here that the tree shows relevant frequencies in a nested-sets formulation (Hoffrage & Gigerenzer, 1998; Hoffrage, Gigerenzer, Krauss, & Martignon, in press). In Figure 3, the nested sets involve the set-inclusion relationship between the numbers of produced gemstones and inspected gemstones at the stages of "production" and "inspection."

Another difference of Experiment 3 from Experiments 1 and 2 lies in the presentation of probability. Recall that in Experiment 2 the prior probabilities for the clear, blurred, and cracked gemstones were given as ½, ¼, and ¼, respectively. Because fractions per se do not always indicate probabilities (e.g., an approximation of π, namely 22/7, is a perfectly legitimate fraction but does not at all qualify as a probabilistic number), probabilities expressed in real numbers would be a more appropriate way to express the uncertainties in the probability instructions. Therefore, in Experiment 3, the following description was adopted as the probability instruction:

> *A factory manufactures artificial gemstones. Each gemstone has a 0.25 chance of being blurred, a 0.25 chance of being cracked, and a 0.5 chance of being clear. An inspection machine removes all cracked gemstones and retains all clear gemstones. However, the machine removes blurred gemstones with the probability of 0.5. What is the chance that a gemstone after the inspection is blurred?"*

In addition, participants in probability groups in Experiment 3 received the above instruction coupled either with the roulette-wheel diagram as in Experiment 2 or with the probability tree in Figure 4. The probability tree in Figure 4 mirrors the frequency tree in Figure 3, except that the numbers are expressed in probabilities (note: not in fractions). Again, the probability tree appears in the nested-sets representation that show the set-inclusion relationship between the probabilities at the stages of "production" and "inspection."

A final change in Experiment 3 from the previous experiments lies in altering the task description for the group that received the roulette-wheel diagram with probability instruction. Unlike Experiments 1 and 2, the description of the roulette-wheel for the

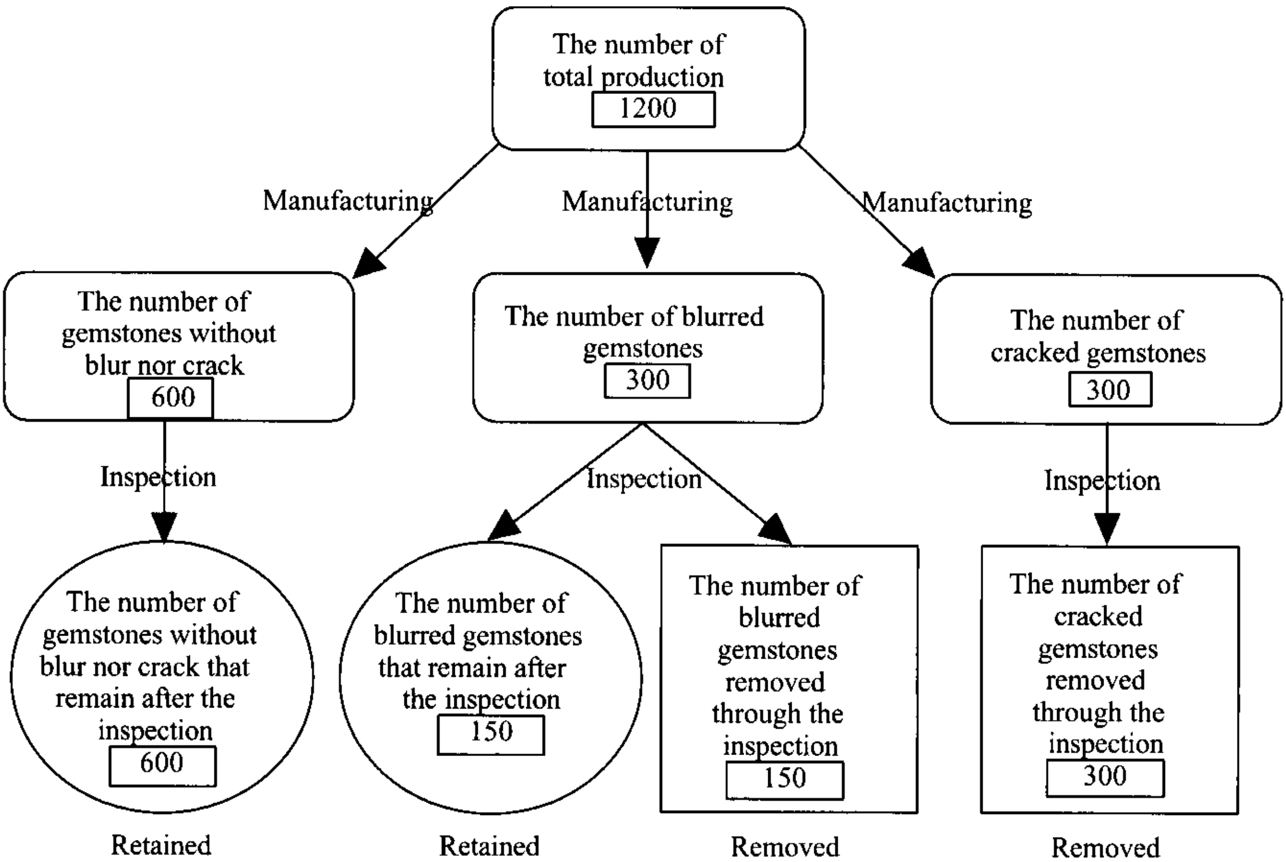

Figure 3. Frequency tree used in Experiment 3.

probability instruction group explicitly stated that the area breakdown in the inner circle and the outer rim represent corresponding *probabilities* for the blurred, cracked, and clear gemstones.

It should be noted here that, because of the introduction of the frequency and probability trees, Experiment 3 does not compare the nested-sets hypothesis against a purely frequency hypothesis, as in Experiments 1 and 2. Instead, Experiment 3 addresses the following questions. First, is it necessary for the frequency instruction to be presented in the nested-sets structure (i.e., the trees) to achieve more normative facilitation in reasoning? Second, would the tree diagrams produce facilitation of normative reasoning comparable to the roulette-wheel diagram?

Independent Variables

The Gemstone Problem was presented either in the frequency instruction (the same verbal description as in Experiment 2, Group 7) or in the aforementioned probability instruction. Another independent variable was the kind of visual aids. The roulette-wheel groups received the same roulette-wheel diagram as in Experiment 2. The frequency-tree group and the probability-tree group received the frequency tree in Figure 3 and the probability tree in Figure 4, respectively, along with the verbal instruction. Thus, $2 \times 2 = 4$ conditions were set up.

Dependent Variable

Participants who received the probability instruction were asked to answer with the intuitive probability. The probability-tree group was instructed to answer either in probability or in an equation that can calculate the answer. Responses from the probability groups were coded correct if the response indicated "0.2" or equivalent answers $\left(\text{such as } \frac{1}{5} \text{ or } \frac{0.125}{0.5 + 0.125}\right)$. Participants in the frequency-version conditions answered with the number of gemstones that passed the inspection and the number of gemstones that are blurred after the inspection. Responses from the frequency-version conditions were coded correct if the response indicated the former and the latter numbers in question as 750 and 150, respectively.

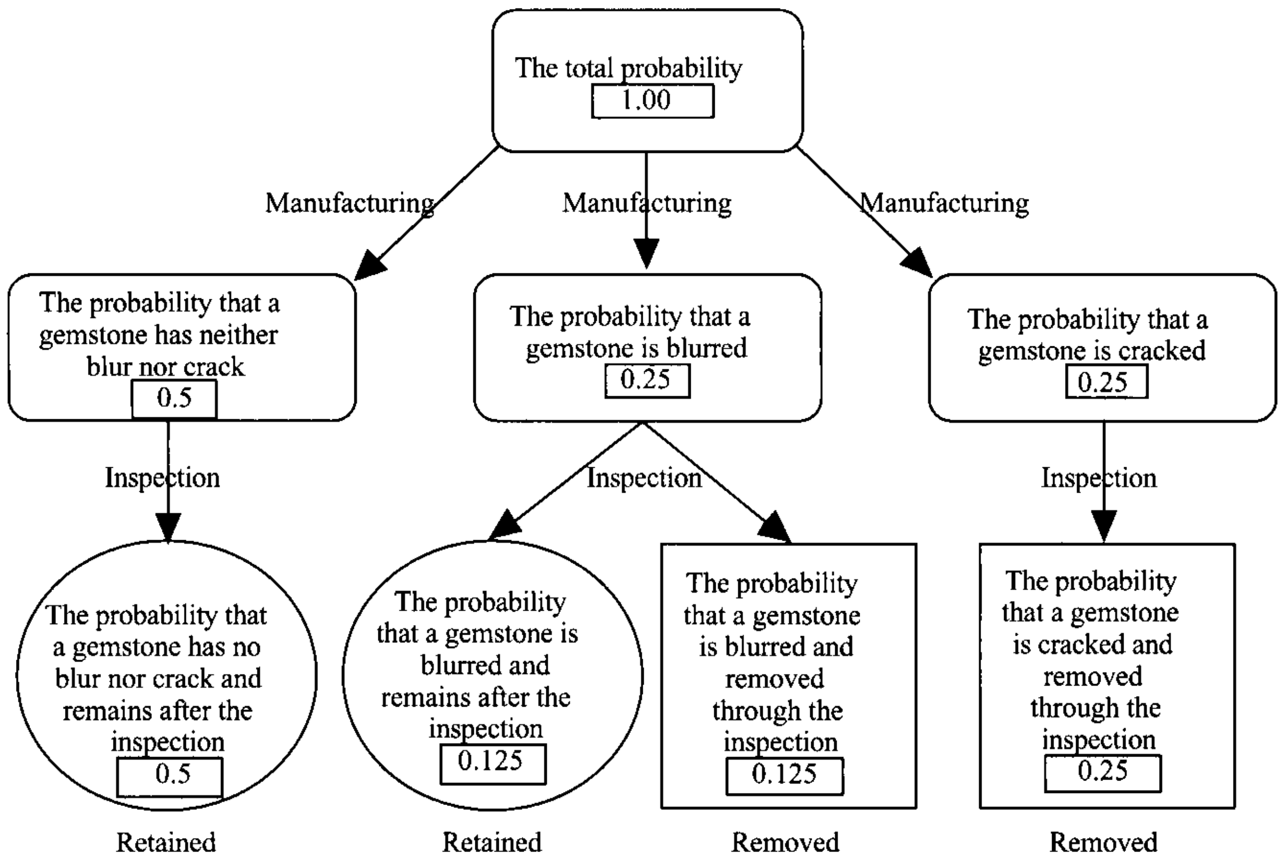

Figure 4. Probability tree used in Experiment 3.

Participants and Procedure

Seven hundred and ninety-eight Japanese undergraduates enrolled in an introductory psychology course participated as a partial fulfillment of the course credit. The rest of the procedure was the same as in Experiments 1 and 2, except that Experiment 3 was run in three separate classrooms. Each participant in every classroom was randomly assigned to one of the four conditions, so that the classes would not produce a confounding factor with the conditions.

Table 3. The Percentages of Correct Responses in Experiment 3

	Frequency Instruction	Probability Instruction
Roulette-Wheel Diagram	Group 9 ($n_9 = 202$) 79.2%	Group 10 ($n_{10} = 197$) 79.7%
Tree Diagram	Group 11 ($n_{11} = 201$) 70.1%	Group 12 ($n_{12} = 198$) 60.1%

Results and Discussion 3

Prior to analyses that are reported in the next paragraph, a preliminary analysis confirmed that the three groups from the different classrooms showed essentially the same trend. Because this indifference bears little implications that may invalidate the following analyses, the issue of indifference among the classes is not developed here any further.

Table 3 shows the percentages of correct responses in each condition. As in Experiments 1 and 2, the roulette-wheel diagram groups (Groups 9 and 10) performed successfully without showing a significant difference in their correct solution rates ($z = -.125$, *ns*). The proportion of correct responses between Group 9 and Group 11 showed a significant difference ($z = 2.145$, $p < .05$), indicating the roulette-wheel diagram's supremacy over the frequency tree. Likewise, Group 10 and Group 11 showed a significant difference ($z = 2.263$, $p < .05$). Therefore, it was demonstrated that the roulette-wheel diagram, regardless of the frequency or the probability instruction, was a more effective facilitator of normative correctness than the frequency tree. A comparison with the probability instruction groups also proves that the roulette-wheel diagram more effectively facilitated normative reasoning than did the probability tree. The correct response rate of Group 10 was significantly greater than Group 12 ($z = 4.383$, $p < .001$). Within the tree conditions, the frequency tree produced more correct responses than the probability tree. The difference between Group 11 and Group 12 showed a reliable difference ($z = 2.132$, $p < .05$). Finally, the effectiveness of tree diagrams was tested against the results from Experiment 2. In the probability instruction groups, the use of the tree diagram showed a significant effect. The performance of Group 12 highly significantly surpassed Group 8 ($z = 7.770$, $p < .001$). However, Group 11 showed only marginal improvement over Group 7 ($z = 1.813$, $p < .07$). Thus, the tree diagrams proved at least marginally more effective in facilitating probabilistic reasoning than verbal instructions alone, although their effectiveness did not compare to the roulette-wheel diagrams.

As for the tree conditions (Groups 11 and 12), a comparison to a previous study should be remarked upon here. In Study 2 of Sedlmeier and Gigerenzer (2001), they administered the probability tree and the frequency tree in their Bayesian reasoning task, and consistent with the current results, the frequency-tree group performed better than the probability-tree group. However, their participants showed higher correct response rates than Groups 11 and 12. It seems possible to raise at least two reasons why Groups 11 and 12 showed such inferior performance. One is the fact that Sedlmeier and Gigerenzer's participants had received an intensive instruction in Bayesian reasoning, whereas the participants in Experiment 3 did not undergo such practice. Another may be due to the nature of the subject pool. The participants in Experiment 3 were recruited mostly from first- and second-year college students majoring in liberal arts. Therefore, Groups 11 and 12 represent a rather mathematically naïve population, and this naïveté may explain why their performance was inferior to the Sedlmeier and Gigerenzer study.

To sum up, the roulette-wheel diagram outperformed the tree diagrams. Although both the roulette-wheel diagram and the tree diagrams qualify as nested-sets representation, the set-inclusion relationship between subsets and supersets are more directly visible in the roulette-wheel diagram. In contrast, faced with the tree diagram, one has to carry out some simple arithmetic to grasp the set-inclusion relationship (such as "300 blurred gemstones consist of 150 that pass the inspection and 150 that are rejected in the inspection: see Figure 3). I would speculate here that the visualization of the set-inclusion relationship was more straightforward in the roulette-wheel diagrams, therefore participants assigned to the roulette-wheel conditions marked better performance than those who worked on the problems with the tree diagrams.

Conclusion

Experiments 1 through 3 provided consistent findings in two respects. First, facilitation of normative statistical reasoning through frequency instruction showed modest success, even with the presence of the tree diagram. Second, the nested-sets instruction outperformed the frequency instruction by using a visualization of the relationship between the prior and the posterior probabilities in the roulette-wheel diagram. In short, frequency did not make the difficulty with the Gemstone Problem "disappear" (Gigerenzer, 1991, 1994), and the roulette-wheel diagram did a superior job. Therefore, although a frequency purist may be compelled to maintain that frequency is the desiderata to achieve normative reasoning, the current results indicate a limitation for this claim. Instead, the results show that obedience to applicable probabilistic norms in reasoning can be better obtained by using problem presentation methods other than frequency instruction. Stibel and Sloman (1999) obtained similar results regarding the "conjunction fallacy" (Tversky & Kahneman, 1982; Tversky & Kahneman, 1983).

Recently, Gigerenzer and colleagues limited the scope of their argument in support of the frequency representation by claiming that it is not frequencies per se, but "natural frequencies" that facilitate proper normative reasoning (Hoffrage & Gigerenzer, 1998; Hoffrage et al., in press). Natural frequencies are contrasted to "normalized frequencies," wherein the number of samples is determined by artificial procedures. In Experiments 2 and 3, the task would classify as an example of normalized frequencies if the Gemstone Problem read:

Three samples of 100 gemstones were taken. The first sample was neither blurred nor cracked. The second sample included 50 blurred gemstones. The third sample contained cracked 50 gemstones. The factory's daily operation manufactures 600, 300, and 300 products of the first, second, and third sample types.

Instead, the frequency instruction of Experiments 2 and 3 read: "*Among the 1200, 300 gemstone are blurred, 300 are cracked, and 600 contain neither,*" indicating that participants in the frequency instruction in Experiments 2 and 3 were presented with natural frequencies. Even so, the results of Experiments 2 and 3 showed that the performance of the natural frequency group (Groups 7 and 11) did not compare to the groups presented with the roulette-wheel diagram (Groups 5, 6, 9, and 10). Hence, it could be concluded that, even though natural frequencies resulted in modest facilitation of proper reasoning, the use of the roulette-wheel diagram produced a superior facilitation of normative reasoning.

A possible rebuttal that may arise from the frequentist school might maintain that the roulette-wheel visualization induces a frequency representation. Such an argument could stem from misunderstanding Mellers and McGraw's (1999) characterization of the relationship between frequency and nested sets that "frequencies...help people visualize nested sets" (p. 419). It must be remarked that Mellres and McGraw's point is not to be confused with saying that nested sets help people visualize frequencies. Figures 1 and 2 may induce a representation compatible with frequency, yet the representation does not show frequencies per se. Rather, the representation symbolizes the relevant nested-sets relations. Moreover, if it were the case that frequency is the desiderata and nested sets suggest frequency implications, Groups 2, 6, and 10 in the current study (diagram-only groups) should have shown comparable or inferior performance to Groups 3, 7, and 11 (natural frequency groups), respectively. The results were exactly the opposite. Therefore, such a counterargument fails to gain support on both logical and empirical grounds.

Finally, I would like to recast the fundamental question: Why did the frequency instruction better facilitate normative reasoning, and why did the nested-sets instruction outperform the frequency instruction? The reader is invited to recall that Gigerenzer (1994), Mellers and McGraw (1999), as well as Sedlmeier and Gigerenzer (2001), commonly called upon visual analogies in characterizing the readiness to comprehend relevant statistical information. Gigerenzer (p. 146) stated: "The frequency format...You can 'see' that the relative frequency is one out of two." Mellers and McGraw noted that frequencies help people *visualize* nested sets, as cited in the previous paragraph. Sedlmeier and Gigerenzer (p. 382) noted: "In the frequency format, one can immediately 'see' the answer." I argue here that the graphical nature of Figures 1 and 2 take advantage of people's automatic visual computation in grasping the relationship between the prior and posterior probabilities. Not surprisingly, visual displays could be the most suitable representation to "visualize" the crucial relationship between pieces of information.

References

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology, 66*, 819–829.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 2, pp. 83–115). John Wiley & Sons Ltd.

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

Goodman, L. A. (1964). Simultaneous confidence intervals for contrasts among multinomial populations. *Annals of Mathematical Statistics, 35*, 716–725.

Griffin, D., & Buehler, R. (1999). Frequency, probability, and prediction: Easy solutions to cognitive illusions? *Cognitive Psychology, 38*, 48–78.

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

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition, 84*, 343–352.

Ichikawa, S. (1989). The role of isomorphic schematic representation in the comprehension of counterintuitive Bayesain problems. *Journal of Mathematical Behavior, 8*, 269–281.

Itoh, Y. (1991). The role of the frequency view of probability in solving the "Problem of Three Prisoners". *Japan Cognitive Science Society Technical Report 19.*

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

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

Mellers, B. A., & McGraw, A. P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage. *Psychological Review, 106*, 417–424.

Mosteller, F. (1965). *Fifty challenging problems in probability with solutions*. Reading, MA: Addison-Wiley.

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

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

Shimojo, S., & Ichikawa, S. (1989). Intuitive reasoning about probability: Theoretical and experimental analyses of the "problem of three prisoners." *Cognition, 32*, 1–24.

Stibel, J. M., & Sloman, S. A. (1999, November). *Do frequency frames make probability judgments more coherent?* Paper presented at the 40th Annual Meeting of the Psychonomic Society, Los Angeles, CA.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*, (pp. 84–98). New York: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293–315.

Yamagishi, K. (1994a). Consistencies and biases in risk perception: I. Anchoring process and response-range effect. *Perceptual and Motor Skills, 79*, 651–656.

Yamagishi, K. (1994b). Consistencies and biases in risk perception: II. Individual subjects' performance. *Perceptual and Motor Skills, 79*, 659–663.

Yamagishi, K. (1997a). Upward versus downward anchoring in frequency judgments of social facts. *Japanese Psychological Research, 39*, 124–129.

Yamagishi, K. (1997b). When a 12.86% mortality is more dangerous than 24.14%: Implications for risk communication. *Applied Cognitive Psychology, 11*, 495–506.

Received May 24, 2002
Revision received August 6, 2002
Accepted September 16, 2002

Kimihiko Yamagishi

Graduate School of Decision Science & Technology
Tokyo Institute of Technology
Ohokayama 2–12–1
Meguro-Ku
Tokyo 152–8552
Japan
Tel.: +81 3 5734 2256
Fax: +81 3 5734 2256
E-mail: kimihiko@ky.hum.titech.ac.jp

