Psychonomic Bulletin & Review
2009, 16 (6), 1065-1070
doi:10.3758/PBR.16.6.1065

Why does the base rate appear to be ignored? The equiprobability hypothesis

Masasi Hattori
Ritsumeikan University, Kyoto, Japan

and

Yutaka Nishida
Osaka University, Osaka, Japan

The base rate fallacy has been considered to result from people's tendency to ignore the base rates given in tasks. In the present article, we note a particular, common structure of the tasks (the imbalanced probability structure) in which the fallacy is often observed. The equiprobability hypothesis explains the mechanism that produces the fallacy. This hypothesis predicts that task material that overrides people's default equiprobability assumption can facilitate normative Bayesian inferences. The results of our two experiments strongly supported this prediction, and none of the alternative theories considered could explain the results.

Since a seminal study by Kahneman and Tversky (1973), a vast number of studies have been concerned with the controversial topic of base rate neglect (for reviews, see, e.g., Barbey & Sloman, 2007a; Koehler, 1996). People are supposed to be insensitive to base rate information when they engage in a task such as the following, modified from Eddy (1982, pp. 251–254; underlines and brackets will be explained later):

> Recently, an incurable disease called X syndrome has begun to be reported. X syndrome is known to show symptoms similar to a cold. Suppose you are a doctor working in a hospital. You are expected to make a judgment about whether or not a patient is infected by X syndrome based on the following information.
>
> The prevalence of X syndrome is 1%. A patient who is infected with X syndrome has an 80% chance of testing positive [having a cough]. However a patient who is not infected with the syndrome has a 9.6% chance of testing positive [having a cough]. Now a patient tests positive [has a cough]. What is the chance that the patient is actually infected with X syndrome? Please answer intuitively. ____%
>
> (Correct Answer: 7.8%)

This is called a base rate task (or a BR task, hereafter). Let $P(H)$ be the base rate (prevalence) of having the disease (i.e., the hypothesis) and $P(D)$ be the probability that a person tests positive (i.e., data). This is a Bayesian inference task to derive the posterior probability of X syndrome, $P(H|D)$, from the true positive (detectability) rate, $P(D|H)$. There is a relationship between the true positive rate and the posterior probability, as follows:

$$P(H|D) = P(D|H) \times \frac{P(H)}{P(D)}. \quad (1)$$

Here, $P(D)$ is expressed as

$$P(D) = P(D|H)P(H) + P(D|\bar{H})P(\bar{H}) = .8 \times .01 + .096 \times .99. \quad (2)$$

Therefore, the correct answer for the task is

$$P(H|D) = .8 \times \frac{.01}{.8 \times .01 + .096 \times .99} \approx .078. \quad (3)$$

From Equation 1, the low base rate, $P(H)$, lowers the posterior probability, $P(H|D)$. However, modal responses in typical experiments are about 80% (see, e.g., Bar-Hillel, 1980), the same rate as the true positive rate. This pervasive and robust phenomenon has been called base rate neglect (or base rate fallacy), because it was believed to demonstrate people's insensitivity to the base rate information.

In considering whether and why the base rate is neglected, we should ask what base rate "neglect" means. As long as a person makes a response, some value must be allocated—at least implicitly—to the base rate, and it is not in this sense ignored. As is obvious from Equation 1, if we suppose $P(H|D) = P(D|H)$, which is a common answer, we can derive $P(H) = P(D)$. This is exactly what base rate neglect implies. If participants assume that the probability of "testing positive" and the probability of "having X syndrome" are equal, they will give the true positive rate as an answer for a posterior probability. In our view, base rates are not ignored, but people make a default equiprobability assumption, $P(H) = P(D)$. Thus, we propose that participants do not neglect the base rate, but that they infer that the posterior and the true positive rates are almost equal by postulating near equality in the

M. Hattori, hat@lt.ritsumei.ac.jp

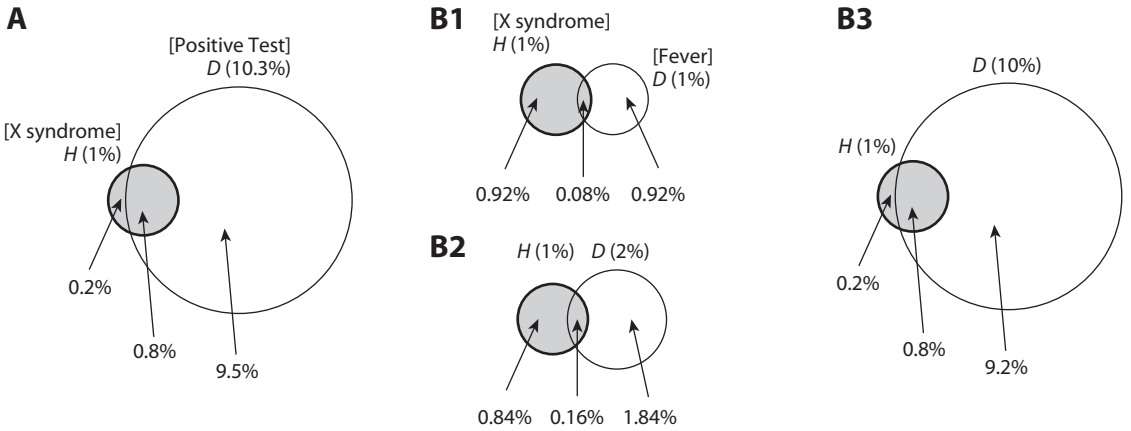


Figure 1. Probabilistic task structures expressed by Euler circles (with correspondence between probability and area size). (A) A task used in Experiment 1. (B1, B2, B3) Tasks used in the F1, F2, and F10 conditions, respectively, in Experiment 2.

marginal probabilities of the two target events that are focused on in the task.

BR tasks share some characteristics that we regard as essential in leading people to the fallacy. The most important feature, we believe, is the fact that the marginal probabilities, $P(H)$ and $P(D)$, differ greatly. Figure 1A (a special Euler diagram) shows the set-size relationship between H and D of the aforementioned task. This is drawn to establish correspondence between each subset probability and each division area of the subsets. Figure 1A shows that (1) the area of $P(D)$ is large in comparison with that of $P(H)$; therefore, (2) although the subset $H\&D$ occupies a large part of set H, it occupies a much smaller part of set D. Consequently, $P(D|H)$ is large, but $P(H|D)$ is small. This is the common distinctive structural feature in the BR tasks and seems to be closely related to the task's difficulty. The structure with a great difference in the marginal probabilities of two target events is called an *imbalanced probability structure* (IPS).

In our view, the BR tasks are difficult in nature because they have an IPS, which conflicts with people's implicit assumption that the two target sets are almost equal in size. In other words, we commit an error in the tasks not because the probabilistic inference is intrinsically difficult, but because two processes—one to recognize the task structure and another to infer the structure on the basis of the equiprobability assumption—compete with each other. According to the equiprobability hypothesis, we predict that (1) people's output responses will roughly conform to the normative Bayesian solution if the default equiprobability assumption is dismissed somehow and the IPS of the task is recognized (Experiment 1), and that (2) people infer more normatively as a task structure becomes closer to the assumed one (Experiment 2).

EXPERIMENT 1
Facilitation of Bayesian Inference

In Experiment 1, we examined the prediction that the recognition of the IPS will facilitate participants' Bayesian reasoning, even in a typical BR task. To encourage participants to give up their equiprobability assumptions and to grasp the task structure, we adopted a technique to utilize their knowledge about the statistical structure of the environment (see, e.g., McKenzie, 2006). Hypothesis H was "being infected with X syndrome" (a fictitious disease), and data D was "having a cough" as a symptom of the disease. X syndrome (H) is a comparatively rare disease [i.e., the base rate, $P(H)$, is low, as in typical BR tasks] and is likely to cause a cough [i.e., $P(D|H)$ is high]. However, having a cough (D) is a common matter in everyday life, so the IPS [i.e., $P(H) \ll P(D)$] can be easily recognized. This is because alternative causes for a cough—including a cold and dust—can be easily brought to mind, and this would promote the recognition that even if the patient has a cough, he or she is not necessarily infected by X syndrome [i.e., $P(H|D)$ is low]. To probe participants' mental representations of the task with IPS, we introduced tasks called "Est-D" and "Est-HD."

Method

Participants. A total of 36 undergraduate students from Ritsumeikan University participated in the experiment as unpaid volunteers. Equal numbers ($n = 18$) of participants were randomly assigned to the cough and PT conditions, described below.

BR task. The tasks were shown at the beginning of the present article. In the positive test (PT) condition (i.e., control), the words "testing positive" (underlined) were used, whereas in the cough condition, the words "having a cough" (in square brackets) were used instead.

Est-D/HD: Tasks for estimating $P(D)$ and $P(H\&D)$. A sample answer sheet for each task is shown in the Appendix. In each task, the total number of people who suffer from X syndrome (H) was set as a reference point for estimation, and participants were asked to graphically estimate the size of subsets: The total number of people who have a cough (or who have tested positive, D) was estimated in the Est-D task; and the total number of people who suffer from X syndrome and who also have a cough (or also have tested positive) was estimated in the Est-HD task. To avoid biasing participants toward any particular size (i.e., toward greater or smaller than H), three options were prepared (see the Appendix).

Procedure. Participants were tested either individually or in groups of 2. The tasks were printed in a booklet, with each task on a

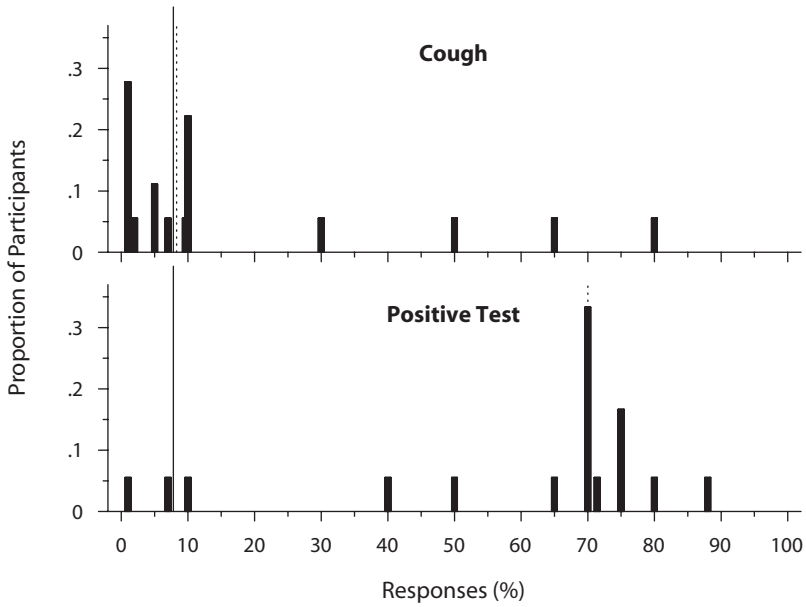


Figure 2. The proportion of participants giving responses plotted against responses (estimated posterior probability, in percentages) in Experiment 1. The upper panel shows the distribution of responses from the task in the cough condition, and the lower panel shows them from the positive test condition. The vertical lines indicate the Bayesian solution, and the dotted lines indicate the median response.

separate page. The booklet consisted of four pages, including a cover sheet with instructions, which stated that participants should read the text carefully before answering the question; they should answer intuitively and need not calculate the answer; there was no time limit and they should perform the tasks at their own pace; and they should answer all three tasks according to page sequence and never in the reverse order. The instructions were read aloud by the experimenter. If participants agreed to participate in the experiment, the experimenter read the text of the BR task. After a participant finished that task, the experiment proceeded to the Est-D and Est-HD tasks.

Results and Discussion

Figure 2 shows the distribution of the estimated $P(H|D)$ in the BR task. As was predicted by the equiprobability hypothesis, the results of the cough condition showed dramatic improvement, whereas the results of the PT condition were very similar to those of previous studies on base rate fallacies. As Figure 3 shows, the data are not normally distributed; therefore, we analyzed them using nonparametric methods. The median answer was 8.3% in the cough condition and 70.0% in the PT condition, and the difference was significant (Mann–Whitney U Test) [$U(18,18) = 48.5, p < .001$].

Answers of 10% or less in the BR task were categorized as correct. This is because 10% as an intuitive answer is close enough to the right answer (7.8%), and there was a considerable gap in the response distribution between 10% and the next smallest answer (i.e., 30% in the cough condition and 40% in the PT condition), as is shown in Figure 2. Proportions of correct answers were 78% (14/18) in the cough condition and 17% (3/18) in the PT condition. The difference was statistically significant [$\chi^2(3, N = 36) = 13.5, p < .001$].

Figure 3 shows a frequency distribution of estimated $P(D)$ values in the Est-D task. This histogram indicates a bimodal distribution. The first mode, located at 0%–2%, is likely to be due to participants who were directly affected by the base rate information (i.e., 1%). They are considered to have assumed equiprobability. The second mode, which peaks at 10%–12%, includes participants who obtained an answer close to the normative Bayesian solution (10.3%) by an intuitive implicit calculation. They are considered to have succeeded in overriding the default

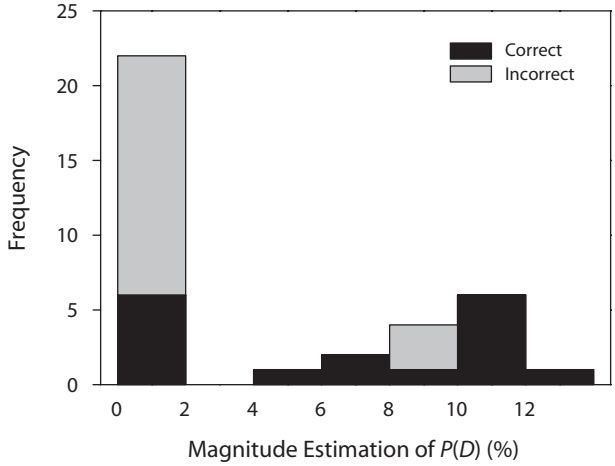


Figure 3. A frequency distribution of the magnitude of *P*(*D*) estimated in the Est-D task in Experiment 1. Estimated values were transformed on a percentage basis. Correct/incorrect indicates responses in the BR task.

(equiprobability) values. By regarding the responses from the first group of participants as incorrect and those from the second group as correct, the relationship between the answers in this task and those of the BR task was analyzed. Although 79% (11/14) of the correct participants in the Est-D task also gave the correct answer in the BR task, only 27% (6/22) of the incorrect participants did so [$\chi^2(3, N = 36) = 9.03, p < .01$]. There was an obvious correlation between answers to the Est-D task and answers to the BR task ($\phi = .50, p < .01$). Similarly, the difference between conditions was significant: 61% (11/18) of participants in the cough condition made correct responses in the Est-D task, whereas only 17% (3/18) did so in the PT condition [$\chi^2(1, N = 36) = 7.48, p < .01$].

A similar result was obtained from the Est-HD task. By considering answers within the range of $0.8 \pm 0.1\%$ as correct, a significant difference in the number of correct responses between conditions was observed: 83% (15/18) in the cough condition versus 50% (9/18) in the PT condition [$\chi^2(1, N = 36) = 4.50, p < .05$]. In the BR task, correct answers were obtained from 54% (13/24) of those who answered correctly on the Est-HD task (24/36), but from 33% (4/12) of those who answered incorrectly on the Est-HD task (12/36), although the difference was not significant [$\chi^2(1, N = 36) = 1.39, p = .23$]. Here again, we can see a correlation—albeit nonsignificant—between the Est-HD task and the BR task ($\phi = .19, p = .27$). In short, it appears that those who comprehend the task structure also succeed in Bayesian inference.

The results confirm the equiprobability hypothesis that the base rate fallacy is caused by the difficulty of recognizing the IPS. The results of the Est-D and DH tasks suggest that the recognition of the IPS, or, more specifically, an (implicit) understanding of $P(H) \ll P(D)$, increases the probability of giving a correct answer in the BR tasks. At the same time, it is significant that merely rewording a reference to a single event (D) makes the fallacy almost disappear. That disproves the charge that people are insensitive to the base rate. Additionally, the clear difference between conditions indicates that existing knowledge can greatly help people to understand the task structure.

EXPERIMENT 2
Imbalance As a Difficulty

According to the equiprobability hypothesis, there is competition between people's default equiprobability assumption and the task structure. If the conflict obstructs people's understanding of the task structure and is the main cause of the base rate fallacy, the degree of conflict will affect their performance. In other words, the degree of imbalance of the task structure will be highly relevant to the difficulty of the task. This is another prediction that was examined in Experiment 2.

Method

Participants. A total of 42 undergraduate students from Ritsumeikan University participated in the experiment as unpaid volunteers. Equal numbers ($n = 14$) were randomly assigned to the three conditions that are described below.

Tasks and Procedure. The method was similar to that in Experiment 1, with the following differences. First, the symptom was neither a cough nor a positive test, but a distinctive "fever."[1] Second, three tasks (i.e., experimental conditions) with different probabilistic information were prepared (see Figures 1B1, 1B2, and 1B3). They shared the same base rate of the disease, $P(H)$, but differed in the probability of the symptom, $P(D)$. As a result, the degree of imbalance between the two target sets, H and D, differed: $P(D)/P(H) = 1$, 2, and 10 in the F1, F2, and F10 conditions (F stands for the factor of imbalance), respectively. Specifically, the base rate probability of the disease, $P(H)$, was "1%," and the posterior probability (i.e., the answer) was "8%" in all conditions. As a consequence, the true positive rate and the false positive rate differed by conditions. The BR task for the F1 condition read as follows (F2 and F10 conditions are listed, respectively, in parentheses).

> Recently, an incurable disease called X syndrome has begun to be reported. Suppose you are a doctor working in a hospital. You are expected to make a judgment about whether or not a patient is infected by X syndrome based on the following information.
>
> The prevalence of X syndrome is 1%. A patient who is infected with X syndrome has an 8% (16%, 80%) chance of having a pathognomonic fever. However, a patient who is not infected with the syndrome has a 0.93% (1.9%, 9.3%) chance of having a pathognomonic fever. Now, a patient has a pathognomonic fever. What is the chance that the patient is actually infected with X syndrome? Please answer intuitively. ____%
>
> (Correct Answer: 8%)

Third, this experiment was administered to participants as a group, and the instructions were not read aloud by the experimenter. Fourth, the Est-D and HD tasks were omitted.

Results and Discussion

As our hypothesis predicted, people were likely to be normative when equiprobability was maintained (F1 condition), and they tended to deviate from Bayesian inference as the structure became distorted (F10 condition). We analyzed the data with nonparametric methods (as were used in Experiment 1) again, because the data were not normally distributed. The median answers were 1.0% in the F1, 3.5% in the F2, and 45.0% in the F10 conditions, and the differences were significant [$\chi^2(3) = 6.01, p < .05$] (Kruskal–Wallis rank sum test), although the difference between the F1 and F10 conditions was marginal [$\chi^2(3) = 5.7, p < .06$] (Scheffé's method).

We adopted the same numerical criterion as in Experiment 1 (i.e., 10% or less), since the data distributions were similar to those in Experiment 1. The proportions of correct answers significantly differed between conditions: 79% (11/14), 64% (9/14), and 29% (4/14) in the F1, F2, and F10 conditions, respectively [$\chi^2(2, N = 42) = 7.58, p < .05$]. The analysis of residuals indicated that F1 was high ($z = 1.98, p < .05$), and that F10 was low ($z = -2.65, p < .01$).

The results indicate that the degree of imbalance in the probability of the two target events determines how difficult it is for people to understand the task structure. This difficulty affects their ability to give a normatively correct response. The results suggest that the accuracy of people's responses will deteriorate only when their attempt to understand the task conflicts with their default equiprobability assumption.

GENERAL DISCUSSION

To support the equiprobability hypothesis fully, we should consider whether the data that seem to confirm it could be explained by other theories. Gigerenzer and Hoffrage (1995) argued that "the mind is tuned to frequency formats, which is the information format humans encountered long before the advent of probability theory" (p. 697). This theory is powerless to explain the results of our present experiment. This would be true of any theory that tried to account for the facilitation of Bayesian inference only in terms of a frequency format and its supposed adaptive advantage from an evolutionary point of view. Any such theory would fail to explain why simple rewording helps people to get the normatively correct answer in the tasks.

Some researchers have attributed the cause of the base rate fallacy to the inverse fallacy (see, e.g., Braine, Connell, Freitag, & O'Brien, 1990; Gavanski & Hui, 1992; Villejoubert & Mandel, 2002). According to this view, the base rate fallacy occurs because people confuse the posterior probability, $P(H|D)$, which is to be estimated, with the true positive rate, $P(D|H)$, given in the task. The equiprobability hypothesis implies the inverse fallacy and explains why it occurs. Any theory based on the inverse fallacy should also account for its origin. Regarding the inverse fallacy as primitive (e.g., Villejoubert & Mandel, 2002) can be an alternative, but this cannot explain without footnotes why people sometimes do not invert conditional probabilities, as was shown in Experiment 1. According to Gavanski and Hui (1992), the factor that makes people answer $P(D|H)$ as an estimation of $P(H|D)$ is the "category–feature" relationship between H and D. In Experiment 1, H was a disease as a category, and D was a symptom as a feature. Therefore, the theory would falsely predict that the inverse fallacy emerges in both of the conditions in Experiment 1. Thus, no inverse fallacy theory appears to fully account for the results of the present study.

The nested sets hypothesis (see, e.g., Sloman, Over, Slovak, & Stibel, 2003) shares a view with the equiprobability hypothesis, laying emphasis on the task structures and their representations in our minds. According to Barbey and Sloman (2007b), "People's ability to estimate the probability of A, given B, in a way that is consistent with Bayes' theorem depends, in part, on the transparency of the structural relations among the set of events of type A, relative to the set of events of type B" (p. 287). Apparently, this hypothesis does not have a view of the IPS. Consequently, the nested sets hypothesis does not explain data from Experiment 2, in which there was no factor that clarified the probabilistic structure of the task.

Internal representation matters. This would be the common claim of the equiprobability hypothesis and the nested sets hypothesis. Barbey and Sloman (2007b) declared, "[The nested sets] hypothesis concerns mental representations" (p. 291), proposing a view that "different external representations (e.g., natural frequencies, chances) map onto the same internal representation" (p. 291). In our view, this would be the most important point of their theory, and we agree with them. In sum, the equiprobability hypothesis parallels the nested sets hypothesis with regard to people's internal structure, but only the equiprobability hypothesis explains why the degree of imbalance affects people's Bayesian inference, as was shown in Experiment 2.

AUTHOR NOTE

The present research was supported by Grant-in-Aid for Scientific Research 19500229 from the Japan Society for the Promotion of Science, awarded to M.H. We thank Yasuaki Kobashi and Yohtaro Takano for their helpful comments on this study. The final version of this article benefited from critical readings by three reviewers. We are especially grateful to David Over for his valuable comments and intelligent advice. Correspondence concerning this article should be addressed to M. Hattori, Department of Psychology, Ritsumeikan University, Kyoto 603-8577, Japan (e-mail: hat@lt.ritsumei.ac.jp).

REFERENCES

Barbey, A. K., & Sloman, S. A. (2007a). Base-rate respect: From ecological rationality to dual processes. *Behavioral & Brain Sciences*, **30**, 241-254. doi:10.1017/S0140525X07001653

Barbey, A. K., & Sloman, S. A. (2007b). Base-rate respect: From statistical formats to cognitive structures. *Behavioral & Brain Sciences*, **30**, 287-292. doi:10.1017/S0140525X07001963

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, **44**, 211-233. doi:10.1016/0001-6918(80)90046-3

Braine, M. D. S., Connell, J., Freitag, J., & O'Brien, D. P. (1990). Is the base rate fallacy an instance of asserting the consequent? In K. J. Gilhooly, M. T. G. Keane, R. H. Logie, & G. Erdos (Eds.), *Lines of thinking: Reflections on the psychology of thought* (Vol. 1, pp. 165-180). Chichester, U.K.: Wiley.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249-267). Cambridge: Cambridge University Press.

Gavanski, I., & Hui, C. (1992). Natural sample spaces and uncertain belief. *Journal of Personality & Social Psychology*, **63**, 766-780. doi:10.1037/0022-3514.63.5.766

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, **102**, 684-704. doi:10.1037/0033-295X.102.4.684

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, **80**, 237-251. doi:10.1037/h0034747

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral & Brain Sciences*, **19**, 1-53.

McKenzie, C. R. M. (2006). Increased sensitivity to differentially diagnostic answers using familiar materials: Implications for confirmation bias. *Memory & Cognition*, **34**, 577-588.

Sloman, S. A., Over, D. E., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior & Human Decision Processes*, **91**, 296-309. doi:10.1016/S0749-5978(03)00021-9

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & Cognition*, **30**, 171-178.

NOTE

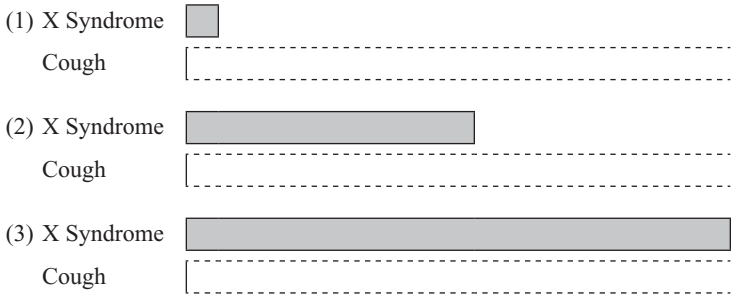
1. The reason for this change was that a more flexible symptom that fits diverse marginal probabilities and bears variable true and false positive rates was desirable. We had to manipulate these rates in order to differentiate the degree of imbalance of the probabilistic structure of the tasks.

(*Continued on next page*)

APPENDIX
Answer Sheet Sample

Est-D Task

The number of patients who are infected with X syndrome is diagramed by the area of a rectangle. How large do you think the number of patients who have a cough would be when you also express it as an area of a rectangle? Please answer intuitively. Rectangles drawn by broken lines below have no right side. You are to draw a single vertical line in it and complete a rectangle with an appropriate size for "the number of patients who have a cough." For your answer, you can choose any one of the three different sized figures from (1), (2), and (3), according to the number of patients you think will have a cough. For example, if you would like to draw a larger rectangle than that of X syndrome then you can use figure (1); for a smaller rectangle use (3); for a similar size rectangle use (2). Note that figures (1), (2), and (3) indicate the same number of people.



Est-HD Task

The number of patients who are infected with X syndrome is diagramed by the area of a rectangle. How large do you think the number of patients who are infected with X syndrome and also who have a cough would be when you also express it as an area of a rectangle as the case of the last task? Please answer intuitively. Rectangles drawn by broken lines below have no right side. You are to draw a single vertical line in it and complete a rectangle with an appropriate size for "the number of patients who are infected with X syndrome and also who have a cough."

(Manuscript received on March 31, 2008;
revision accepted for publication July 20, 2009.)