# Chapter 2: Applying natural frequencies to teach medical students how to interpret diagnostic test results

*"Good medicine does not consist in the indiscriminate application of laboratory examinations to a patient, but rather in having so clear a comprehension of the probabilities of a case as to know what tests may be of value … it should be the duty of every hospital to see that no house officer receives his diploma unless he has demonstrated … a knowledge of how to use the results in the study of his patient."*

(Peabody, 1922; cited from Casscells et al., 1978, p. 1000)

This chapter is concerned with the ability of (future) medical experts to draw diagnostic inferences. Diagnosing diseases is an important task most physicians face every day, and often an array of tests and tools is at hand to assist in gathering evidence for and against the diagnosis in question. Interpreting the results of such tests leads to an updating of one's prior belief in the presence of the disease: What is the probability, for instance, that a woman has breast cancer when she has a positive mammogram? How certain is it that an unborn child has Down's syndrome, given a suspicious result from an ultrasound scan? As medical decisions on further diagnostic procedures and treatments are based on such probability judgments (Eddy, 1982), it is important for doctors to be able to estimate these probabilities accurately.

As illustrated in the introductory chapter, one needs to combine information on the prevalence of the disease with information on the accuracy of the test, that is, its sensitivity and specificity. But, as also said before, previous research has shown that not only lay people, but also physicians have problems in drawing such so-called Bayesian inferences (e.g., Eddy, 1982). These findings emphasize the need to include courses on statistical reasoning in general and Bayesian reasoning in particular in medical education (Gigerenzer, 2000; Gigerenzer & Edwards, in press).

Following this suggestion, imagine a professor preparing a class in medical school on the topic of prenatal diagnosis. She wants to give the students an overview of the several methods available for detecting Down's syndrome and to instruct them on how to judge the predictive value of these tests. In the introduction to the class, the professor confronts the students with the following problem (statistical information adapted from Howe, Gornall, Wellesley, Boyle, & Barber, 2000; Snijders, Noble, Sebire, Souaka, & Nicolaides, 1998):

*The probability that a pregnant woman gives birth to a child who has Down's syndrome is 0.15%. If a woman is pregnant with a child who has Down's syndrome, the probability that the ultrasound test on nuchal translucency shows positive is 80%. If the woman is pregnant with a child who does not have Down's syndrome, the probability that the ultrasound test still shows positive is 8%. What is the probability that a woman is pregnant with a child who has Down's syndrome, given a positive ultrasound test?*

The problem has the same structure as the mammography problem in the introductory chapter. The correct solution is 1.5% and can again be obtained by inserting the statistical information on the prevalence of the disease together with the sensitivity and false-alarm rate of the test into Bayes' rule. This being said, a straightforward approach for the professor would be to teach Bayesian reasoning by training her students how to insert the statistical information into Bayes' rule, hoping that this instruction would overcome their difficulties with the task of probability updating. However, previous training studies using this rule-training approach have not been very effective. According to a recent review (Sedlmeier, 1999), only a few training studies that addressed Bayesian reasoning are reported in the literature, and neither of the approaches applied, be it rule-training or the provision of corrective feedback, yielded a substantial training effect.

An alternative method for overcoming these difficulties can be inferred from the above-mentioned findings of Gigerenzer and Hoffrage (Gigerenzer & Hoffrage, 1995; Hoffrage & Gigerenzer, 1998) on the facilitating effect of natural frequencies: Rather than instructing her students on how to insert the statistical information into Bayes' rule, our professor could train them in representing statistical information in terms of natural frequencies. As statistical information in medical textbooks, newspapers, and other media is most often displayed in a probability or percentage format, the training should enable participants to translate probabilities into natural frequencies. It is important to note the difference in the use of the tool of natural frequencies in the training approach presented here, compared to the studies presented in the introductory chapter (Hoffrage & Gigerenzer 1998, in press): In the latter, performance was increased when natural frequencies were presented *instead* of probabilities. Here, the idea is to give participants statistical information in terms of probabilities, but to instruct them to *translate* them into natural frequencies in the process of solving the task.

Sedlmeier and Gigerenzer (Sedlmeier, 1997; Sedlmeier & Gigerenzer, 2001) were the first to develop a tutorial based on the representation-learning idea. They designed a 2-hour computerized tutorial in which participants – here: university students from differing fields – learned to solve Bayesian tasks in individual training sessions by translating probabilities into natural frequencies. For comparison, participants in another group received the traditional rule training. Sedlmeier and Gigerenzer could show (2001) that the proportion of correct answers immediately after the training was clearly higher when participants had learned to represent probabilities as natural frequencies (75% correct answers when the frequencies were presented in a grid and 90% when they were presented in a tree), as opposed to inserting them into Bayes' rule (60% correct answers). Moreover, the learning effect for the representation-training groups was far more stable, as performance in these groups remained on this high level even in a post-test given 5 weeks after training, whereas performance in the rule-training group dropped to 20%.

Let us come back to the medical context and the question how medical students should be instructed to draw diagnostic inferences. Obviously, the representation learning approach in form of the computerized tutorial by Sedlmeier and Gigerenzer (2001) is a very promising method. But unfortunately, it cannot yet be directly implemented in the typical educational setting of German medical schools. The reason is that computer tutorials are not yet part of the regular curricula in most German medical schools, especially not in statistics courses. If a tutorial on Bayesian reasoning should take place today, it should rather have the following features: It has to take place in the classroom with a group of students, because single training sessions are not possible. The material should be presented on overhead slides or the blackboard; individual training computers are not available in the classes. There is usually not enough time for the teacher to provide individual feedback to the students. Moreover, although the medical school that I cooperated with (Institute for Medical Genetics, Free University of Berlin) was very interested in the tutorial, it could only offer a one-hour slot for such a tutorial in the current curriculum. All this implies that, to match these particular demands, the tutorial had to be modified with respect to group size, media, feedback and duration. Therefore, a one-hour classroom tutorial on Bayesian reasoning based on the representation-learning approach outlined above was developed and tested in a human genetics course for medical students. To evaluate the relative effectiveness of the modified tutorial, a second treatment condition was included in the study in which participants received traditional rule training.

# Study 1: Evaluation of a classroom tutorial on Bayesian reasoning for medical students

What are the predictions on the effectiveness of the two variants of the classroom tutorial? Sedlmeier and Gigerenzer (2001) did not discuss the question of how the specific instructional setting contributes to the performance, or how performance would be affected by changes in the instructional setting. However, it seems plausible to assume that the modifications in the classroom tutorial are less optimal with respect to interactivity and individual activity than flexible computerized tutoring systems and therefore lead to lower absolute performance rates, compared to the computerized tutorial. Because the classroom tutorial differs in several aspects from the computerized tutorial, it will not be possible to attribute differences in the absolute performance rates between the classroom and the computerized tutorial to one specific feature in the classroom setting. However, the modifications affect both treatment conditions (representation and rule learning) alike. Therefore, the following prediction can be maintained also for the classroom setting: Performance after the representation-learning training should be better than after the rule-learning training.

## Method

### *Participants*

Participants were 208 medical students in their second and third semesters, in an obligatory all-day course on human genetics at the Free University of Berlin. The tutorials took place at the end of the seminar day, due to external constraints imposed by the organizing institution, and students were at this point divided into groups with an average of 16 participants.

### *The classroom tutorials*

Because the tutorials had to be easy to implement in a classroom setting, the only media used were overhead slides and handouts. In addition, the tutorial contents had to meet the specific needs of medical students at the beginning of their studies: First, the illustrating examples of the tutorial were chosen to relate to the contents of the course in session (here: human genetics); second, a short input section on technical terms such as prevalence and

sensitivity was added to provide background knowledge that would facilitate subsequent autonomous use of statistical literature.

The representation-learning tutorial consisted of four main parts: In the introductory first part (15 minutes), the Down's syndrome problem was introduced as an initial probability-updating problem and briefly discussed. Participants then had to record their individual estimations, which were collected and served as the pre-test. The second part (10 minutes) was a short input section on the terms prevalence, sensitivity, specificity, false-alarm rate, and positive predictive value. These concepts were illustrated with a 2 x 2 table (disease absent/present vs. test positive/negative). In the third part of the tutorial (15 minutes), participants were instructed on how to solve the initial problem. They learned how to translate the probability information into natural frequencies, graphically aided by completing a frequency tree similar to Figure 1.1 (see Appendix A for materials). The translation steps were as follows:

- Select a population (e.g., 10,000 people) and use the base rate to determine how many of the population have the disease (in the case of Down's syndrome, 0.15% of 10,000 unborn children is 15 children).
- Take that result (15 children) and use the test's sensitivity to determine how many have both the disease and a positive test (80% of 15 is 12 children).
- Take the remaining number of people who do not have the disease (9,985 children) and use the test's false positive rate to determine how many do not have the disease, but still test positive (8% of 9,985 is about 799 children).
- Compare the number obtained in Step 2 with the sum of those obtained in Steps 2 and 3 to determine how many people with a positive test actually have the disease (12 out of 811 children).

In the fourth part of the tutorial (20 minutes), students were required to solve three additional text problems by themselves, while guidance by the tutor gradually declined. Here, learning-by-doing was encouraged to foster the acquisition of procedural knowledge, an essential ingredient of successful instruction (Anderson, Corbett, Koedinger, & Pelletier, 1995; Sedlmeier, 1999). Subsequently, feedback was given by discussing individual solutions and answering any remaining questions.

In the rule-training tutorial, contents, sequence, and duration were exactly the same as in the representation training, the only difference being in the contents of part three, the instruction part. Here, participants were first extensively introduced to Bayes' rule and its components. They were then shown how to insert the probabilities into the formula,

illustrated by the completion of the formula on an overhead slide. Please note, however, that there was no difference between the two tutorials with regard to the duration of the instruction part.

### *Design*

Of 208 participants, 109 received representation-training (7 groups) and 99 rule-training (6 groups). The groups were randomly assigned to one of the two training conditions, and also the four tutors of the study were randomly assigned to the training groups.

The evaluation study had a pre- and post-test design; both tests consisted of one Bayesian text problem with information given in the probability format. The pre-test consisted of the Down's problem already mentioned and was given at the beginning of the training session. The post-test also consisted of one text problem (one half of the participants in each training group received a problem on mammography, the other half a genetic test for diabetes) and was administered 2 months following training to assess the training's long-term effects. There were two ways of administering the post-test: Either participants completed the post-test in the final 15 minutes of another seminar that some of the training groups were required to attend, or they received the post-test by mail (due to different curricula, not all training groups were able to meet again 2 months later). To increase the return rate, these participants were paid 5 Euro for sending back the questionnaire. In the instructions, participants were urged not to use notes from the tutorial. It was explained that if they did, the questionnaires would be of no use for the study, and it was pointed out that payment would occur independent of the accuracy of their solution.

### *Assessing performance*

In both the pre- and the post-test, participants were requested to record their estimate of the positive predictive value and to note how they arrived at their solution. Following the methods used in previous studies (Gigerenzer & Hoffrage, 1995), a solution was classified as Bayesian, and therefore correct, when (a) the estimate was in a range of plus/minus 1% of the correct solution and (b) the notes indicated that the estimate actually resulted from a Bayesian approach (with natural frequencies, Bayes' rule or a shortcut thereof, see Gigerenzer & Hoffrage, 1995). This rather conservative double scoring criterion was used to obtain a more valid measure of the effects of the classroom tutorial, that is, one that can clearly distinguish between the correct strategy taught in the tutorial on the one hand and incorrect strategies or guessing on the other hand.

## Results

### *Pre-test*

Of the 109 participants in the representation training who completed the pre-test, one gave a correct solution (1%). Of the 99 participants in the rule training, three gave a correct answer (3%). These results indicate that prior to training, participants in both groups had few skills for solving this type of task correctly. Compared to other studies (Gigerenzer & Hoffrage, 1995; Hoffrage & Gigerenzer, 1998), the rate of 1 – 3% correct solutions is rather low, indicating a relatively low level of statistical education in this sample.

### *Post-test*

As described above, there were two methods for collecting data for the post-test: 128 students received the post-test by mail (72 representation, 56 rule), whereas 78 students[2] worked on the test at the end of another seminar in the institute (35 representation, 43 rule). Fifty-seven mailed questionnaires (31 representation, 26 rule) were returned, which corresponds to a return rate of 45%. There were no substantial differences in the return rates for the representation group compared to the rule group (43% and 46%) and for the two post-test problems (mammography problem 46%, diabetes problem 43%). Overall, post-test data from 66 participants in the representation-learning group and 69 in the rule-learning group were considered in the subsequent analysis. I employed Chi-square ($\chi^2$) tests to examine differences in the proportion of correct and incorrect answers.

How many participants succeeded in solving a Bayesian text problem two months after training? As Table 2.1 shows, 47% (31 of 66) of the students who learned to translate probabilities into natural frequencies gave a correct answer on the post-test, compared to 16% (11 of 69) of the students in the rule-training group ($\chi^2$ (1, 135) = 15.15, $p < .001$, $\phi = .34$).

Table 2.1
Absolute number of correct inferences for the two trainings and the two post-test conditions

| Post-test setting | Rule training | Representation training | Total |
|---|---|---|---|
| Course | 2 (43) | 13 (35) | 15 (78) |
| Mail | 9 (26) | 18 (31) | 27 (57) |
| Total | 11 (69) | 31 (66) | 42 (135) |

*Note*. Numbers in parentheses are the total number, correct and incorrect, of inferences per cell.

[2] Originally, 80 students received the test in the seminar, but two did not want to do the test and dropped out.

The phi coefficient $\phi$ was calculated as a measure of effect size. According to Cohen (1988), $\phi$ values of .10, .30, and .50 correspond to "small", "medium", and "large" effect sizes. By this classification, the effect of the training condition on performance in the post-test was "medium". There was no difference in performance for the two post-test tasks ("mammography": 20 correct of 68, "diabetes": 22 correct of 67). Figure 2.1 illustrates the effect of the two training programs: Students could profit from both trainings, but the learning effect that could be observed two months after the training was much stronger in the representation-training group.

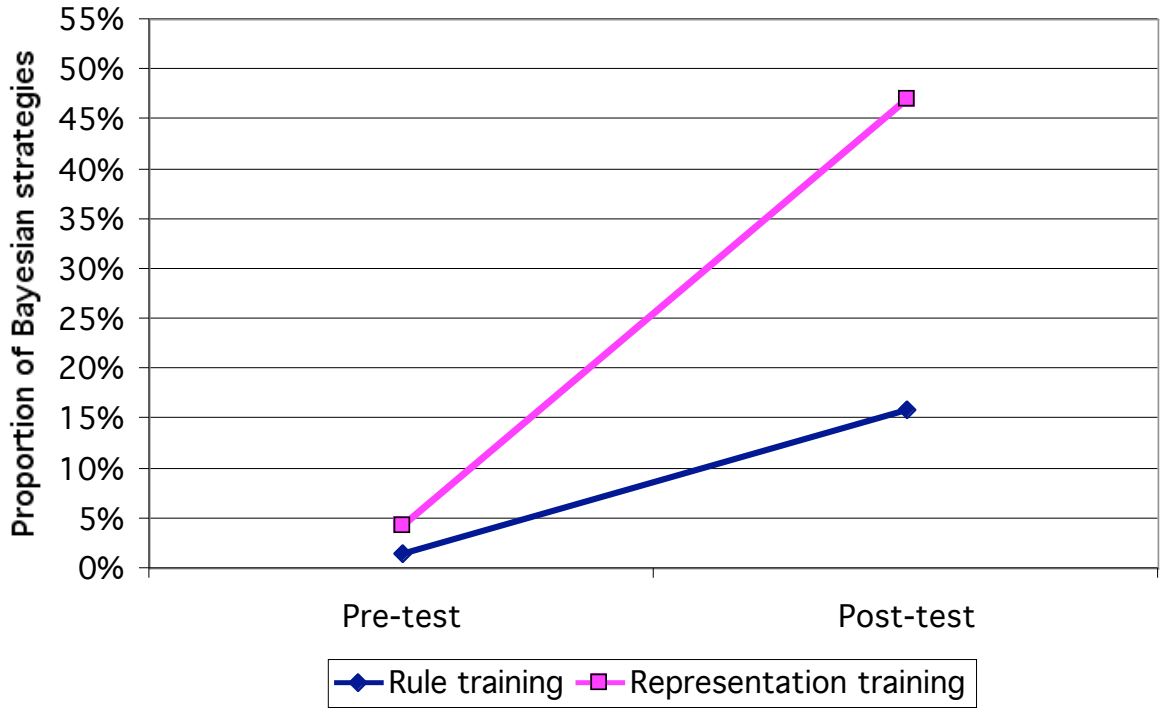


Figure 2.1 Medical students' percentages of correct inferences for the two training conditions before and two months after the training.

There was an unexpectedly strong effect of the post-test setting on performance ($\chi^2$ (1, 135) = 12.17, $p < .001$, $\phi = .30$). In both post-test conditions, students profited more from the representation-training than from the rule-training (Table 2.1). However, the percentage of accurate responses was significantly higher for the mailed questionnaires (27 correct of 57, 47%) than for the questionnaires completed in the course (15 correct of 78, 19%).

What strategies did participants use when solving the task? Four categories were used to classify the identifiable strategies: Bayesian correct (Bayesian approach and correct solution, criterion see above), Bayesian incorrect (starting with the Bayesian approach but arriving at an incorrect solution, for example, frequency tree correct, but incorrect division of

values), non-Bayesian (strategy clearly different from the Bayesian approach and incorrect solution, for example, base rate divided by sensitivity), and guessing. Two additional categories were used for unidentifiable and for missing strategy information. Table 2.2 shows the differences in strategy application between the two training groups: In addition to the previously mentioned higher percentage of correct Bayesian approaches in the representation group, the groups differed mainly in the proportion of participants who produced unidentifiable strategy notes or no notes at all, which was clearly higher for the rule-learning group (24 of 69 answers) than for the representation-learning group (6 of 66 answers). There was also a variety of non-Bayesian strategies that participants from both groups used, but none of these non-Bayesian strategies was applied by more than two participants.

Table 2.2
Medical student's strategies in the post-test dependent on the two training conditions

| | Rule training | | Representation training | |
|---|---|---|---|---|
| Strategy | *N* | % [a] | *N* | % [b] |
| Bayesian correct | 11 | 17% | 31 | 47% |
| Bayesian incorrect | 16 | 24% | 19 | 29% |
| Non-Bayesian | 12 | 18% | 9 | 14% |
| Guessing | 6 | 9% | 1 | 2% |
| Not identified | 10 | 15% | 3 | 5% |
| Missing | 14 | 21% | 3 | 5% |

*Note.* [a] Percentages refer to $N = 69$ participants in rule training. [b] Percentages refer to $N = 66$ participants in representation training.

Although the use of non-Bayesian strategies was less frequent in the representation group than in the rule group, the percentage of Bayesian-incorrect approaches was higher in the representation group. The errors participants produced here provide information that can be used to improve the tutorial further. The most frequent error in the rule-training group was that Bayes' rule was memorized incorrectly and, therefore, its application led to incorrect results (n = 10). In the representation group, the two most common errors were an incorrect setup of the frequency tree (n = 7) and, after depicting a correct tree, an incorrect integration of numbers; for example, five participants used the probability of joint occurrence of disease and positive test result only as an estimate without further division.

***Students' comments on the tutorial***

For three representation- and four rule-training groups, the tutors distributed questionnaires at the end of the tutorial, asking for any remaining open questions, comments, or critique. Altogether, 75 questionnaires were returned with 159 comments. Approximately one half of the comments consisted of general remarks, with positive statements (69) such as

"the tutorial was interesting and explained comprehensively" being much more frequent than negative statements (6). This implies that the majority of students in both training conditions evaluated the tutorial positively. Additional analysis showed that the participants' general attitude was independent of their performance. Remaining open questions were addressed by 26 comments, mainly from participants in the rule-learning group (18 of 26) who had queries concerning the origin of the statistical information in the examples and the meaning of the components of Bayes' rule, such as sensitivity. Which aspects of the tutorial should be improved, according to the participants? Equally divided between the training groups, 19 of 50 remarks stated that more than one hour should be devoted to this topic, with the consequence that more time could be spent discussing the examples. The second most frequently mentioned criticism was that the time of day for the training was not optimally chosen, namely, at the end of a long day when the participants' attention and motivation were at a relatively low level.

**Discussion**

The present evaluation study shows the differential effectiveness of the two approaches to teaching Bayesian reasoning. Whereas both improved performance compared to the pre-test, almost three times as many students were able to profit from the representation-learning tutorial than from the rule-learning tutorial two months after the training. As hypothesized in the introduction of this chapter, the absolute level of performance in both conditions of the classroom tutorial was lower than that of the computerized tutorial by Sedlmeier and Gigerenzer (2001), which is probably due to the less optimal learning conditions in the classroom tutorial. However, the relative effectiveness of the representation-learning approach compared to the rule-learning approach was clearly higher in both instructional contexts, classroom and computer tutorial.

I would like to discuss two issues that seem especially relevant for future applications of the classroom tutorial. First, there was the rather unexpected result that the two post-test conditions lead to differences in performance rates. Due to organizational constraints, not all participants could complete the post-test in the seminar, and some had to have a questionnaire mailed home; the proportion of correct responses was higher for the latter group. The level of performance was obviously influenced by the post-test setting, and I would like to mention three reasons that could possibly account for this effect. One possibility is self-selection, in the sense that students who could not solve the task correctly might have been less likely to

return the questionnaire, thereby inflating the proportion of correct responses among those that were returned. A second reason could be cheating, that is, some students may have used the opportunity to check their notes, ignoring the request not to do so. Finally, it could also be possible that actual performance decreased in the course condition, due to the unfortunately noisy, overcrowded setting in which the post-test had to be collected (a high proportion of post-tests with missing strategy information in these groups supports this interpretation). It cannot be determined, at this point, which interpretation is factual, and future research will have to clarify to what extent the post-test setting can influence assessment of training effectiveness[3].

Second, another important point is revealed in the strategy analysis: Between a quarter and a third of the students began with the correct approach but then failed in the process of solving the task. As a reaction to the most common errors, more time should be spent in the tutorial explaining Bayes' rule, the transfer of probability information into a frequency tree, and the final integration of frequency information. This could either occur by expanding explanations in the instruction phase or, more promisingly, by extending the learning-by-doing phase at the end. Both modifications imply that more than one tutorial hour would be desirable, which is in line with the request made by the participants for more time and better timing. Thus, the one-hour tutorial tested here can be considered as a minimal version that yielded good results, but an extended version of 75 minutes or 2 hours could be even more effective.

However, even without extension, the representation-learning approach was clearly more effective in teaching Bayesian reasoning to medical students than the rule-learning approach. Two main conclusions can be drawn from this study. The first is that the representation-learning approach is more beneficial than the rule-learning approach, and this effect holds for different instructional settings, be it the classic classroom situation or modern computer-based tutorials. Moreover, a recent study by Krauss and colleagues shows (Krauss, Martignon, Hoffrage, & Gigerenzer, 2002) that representation-learning is also efficient in instructing medical students on how to deal with more complex diagnostic problems that invoke data from more than one observation, for example, two medical tests in a row. The second conclusion is that the computerized tutorial used by Sedlmeier and Gigerenzer (2001) is the more effective instructional setting, compared to the classroom tutorial presented here.

[3] To add another method of assessing post-test performance to the list, the tasks can also be included in regular examinations. When a professor from a different department applied the representation-learning classroom tutorial (without rule training as a control group), she found that, after 2 months, 78% of all students answered a probability task accurately.

However, as long as computerized tutorial are not yet available for medical students in their regular classes, the classroom tutorial is a satisfying and easy-to-implement alternative.