Chapter 3: How exactly do natural frequencies facilitate the interpretation of statistical information? Implications for the application of natural frequencies outside the text problem paradigm

The preceding chapter dealt with the question of how to instruct medical students, that is, future medical *experts*, to draw correct diagnostic inferences. The following chapters will focus on a different question, namely how to educate medical *lay people* about the uncertainties and risks associated with diagnostic tests.

Based on the beneficial effect that the natural frequency format has on solving diagnostic inference tasks, it seems straightforward to recommend that also in the communication of risks from medical experts to lay people, statistical information should be represented in terms of natural frequencies to foster comprehension. But how can we be sure whether the tool that worked well in the context of Bayesian text problems will have a similarly beneficial effect in the applied context of understanding medical risk communication? To answer this question, let us consider the differences between the task to solve a text problem and the task to understand risk communication more closely.

People who work on a Bayesian inference problem, such as the medical students in the previous study, have the primary task to *infer information that is not given* from the information given by means of computation. For the addressees of medical risk communication, the primary task is to *understand the given information* correctly (Wright, 1999a), not necessarily to perform computations on it. For instance, the predictive value of a positive mammogram should ideally already be mentioned in a good mammography pamphlet, so that the reader does not have to infer this value from other information. Thus, natural frequencies should have more facilitating features than the one that reduces the number of necessary computational steps in Bayesian inference problems in order to facilitate comprehension outside the context of text problems. And in fact, natural frequencies have several of these features.

First, natural frequencies automatically specify the reference class of the statistical information. The reference class is the class of objects or events to which a probability or frequency refers (Gigerenzer, 2002), for instance “8 out of 10 *women who have breast cancer* will be receive a positive result in mammography screening”. As we will see in the following chapters, missing or ambiguous reference classes are a major cause of misunderstandings of

statistical information (Gigerenzer, 2002; Gigerenzer & Edwards, in press; Phillips, Glendon, & Knight, 1999). A classical non-medical example is the rain probability that is often mentioned in weather forecasts. The statement "There is a 30% chance of rain tomorrow" is interpreted quiet differently by people. Most frequent are the following three interpretations: It will rain in 30% of the area, or it will rain in 30% of the time, or it will rain on 30% of the days like tomorrow (Gigerenzer, Hertwig, van den Broek, Fasolo, & Katsikopoulos, in press). The reason for the ambiguity of the statement is that it contains a probability about a singular event (it will either rain or not rain tomorrow), and such single-event probabilities by definition leave open the class of events to which the probability refers. To avoid confusion, single-event probabilities should be combined with explicit information about the intended reference class, or they should be replaced with natural frequencies (Gigerenzer, 2002; Gigerenzer & Edwards, in press; Gigerenzer et al., in press). Statements with relative frequencies, as the name already indicates, typically include the reference class (e.g., "80% of the *women who have breast cancer* will ... ") – but not automatically. Percentage statements can often be found without any indication of the reference class (e.g., "Mammography screening reduces breast cancer mortality by 30%").

Second, natural frequencies are cardinal numbers, while percentages and probabilities are fractions. Cardinal numbers are easier to visualize than fractions and might thus facilitate the construction of correct internal representations of the information (Gigerenzer & Hoffrage, 1999; Slovic, Monahan, & MacGregor, 2000). Moreover, Gigerenzer and Hoffrage (1995) argued that dealing with cardinal numbers is computationally less demanding than dealing with fractions. Remember that this aspect was part of their computational explanation (see Chapter 1) for the facilitating effect of natural frequencies. They argued that the computational demand of a Bayesian inference task has two components, the number of computations (multiplication, addition, or division) that need to be performed to arrive at the correct solution, *and* the type of numbers that these computations are performed on (absolute frequencies vs. fractions). The latter claim can be tested by investigating how people solve a variant of the Bayesian inference problem introduced so far, namely, Bayesian problems in the so-called *short information menu*. Short-menu tasks can be constructed for all statistical formats, and as will be shown below, the number of computations in the short menu is the same for both natural frequency and probability formats. According to Gigerenzer and Hoffrage (1995), short-menu tasks with natural frequencies (subsequently called *short frequency* tasks) should lead to better performance than short-menu tasks with probabilities or relative frequencies (*short probability* tasks and *short relative frequency* tasks).

However, the evidence base for this argument is inconclusive. While three studies found an advantage of natural frequencies over probabilities in the short information menu (Gigerenzer & Hoffrage, 1995; Hoffrage et al., in press; Mellers & McGraw, 1999, Study 1), two studies did not (Fiedler et al., 2000; Mellers & McGraw, 1999, Study 2).

The goal of the present chapter is to clarify why these seemingly inconsistent results were observed and whether there are boundary factors for the facilitating effect of natural frequencies that could affect the effectiveness of this tool in the risk communication context. First I introduce the short information menu and the evidence for format effects in these tasks obtained so far. I then focus on the studies of Gigerenzer and Hoffrage (1995) and Fiedler et al. (2000) and propose several hypotheses about what could have caused the contradictory results. Finally, I report two experiments that tested four of the hypotheses.

At this point, I would like to forewarn those readers of this dissertation who are primarily interested in medical risk communication, and less so in Bayesian text problems: The following analysis will be very detailed and, unlike the rest of the dissertation, will not include references to applied research but draw exclusively on the literature on Bayesian reasoning. However, you will find a discussion of some applied implications of this analysis at the end of the chapter.

What is a "short information menu"?

The term *information menu* refers to "the manner in which information is segmented into pieces within any format" (Gigerenzer & Hoffrage, 1995, p. 687). The Bayesian inference problems in Chapter 1 and 2 had a *standard information menu* that segments the information into three pieces, $p(H)$, $p(D|H)$, and $p(D|\neg H)$. The *short information menu*, as introduced by Gigerenzer and Hoffrage (1995), segments the information into two pieces of information: $p(H \& D)$, and either $p(D)$ or $p(\neg H \& D)$ (i.e., the number of correct positives, and either the number of all positives or the number of false positives)[4]. Just like standard menus, short menus can be constructed for all statistical formats. Table 3.1 shows the text of the short-menu mammography problem in three different formats, as used in Gigerenzer and Hoffrage (1995; they used the combination $p(H \& D)$ and $p(D)$).

[4] Please note that when I speak of short menu tasks in this chapter, I include not only those tasks that contain *exactly* these two pieces of information, but all tasks that contain *at least* these two pieces of information. For example, in the short menu tasks of Mellers and McGraw (1999), three pieces of information were mentioned: the correct positives, the false positives, and the base rate. This is still a short menu task, because participants can derive the positive predictive value directly in the way specified below; the base rate here is not relevant for the computation (although such seemingly irrelevant information could still affect performance; see below).

Table 3.1
Three short menu versions of the mammography problem (adapted from Gigerenzer & Hoffrage, 1995)

Format	Problem text – Short menu
Natural frequency	103 of every 1,000 women at age forty who participate in routine screening will have a positive mammogram. 8 of every 1,000 women at age forty who participate in routine screening have breast cancer *and* a positive mammogram. Here is a new representative sample of women in this age group who had a positive mammogram in a routine screening. How many of these women do you expect actually to have breast cancer? ___ out of ___
Relative frequency	10.3% of women at age forty who participate in routine screening will have a positive mammogram. 0.8% of women at age forty who participate in routine screening have breast cancer and a positive mammogram. A woman in this age group had a positive mammogram in a routine screening. What is the probability that she actually has breast cancer? ___%
Single-event probability	The probability that a woman at age forty who participates in routine screening will have a positive mammogram is 10.3%. The probability that a woman at age forty who participates in routine screening will have breast cancer and a positive mammogram is 0.8%. A woman in this age group had a positive mammogram in a routine screening. What is the probability that she actually has breast cancer? ____%

Gigerenzer and Hoffrage (1995, Study 1) found that performance in the short menus was, overall, better than in the standard menus, with only a small difference between the two frequency menus (standard: 46% correct, short: 50% correct) and a larger difference between the two probability menus (standard: 16% correct, short: 28% correct). How can the differential impact of the short menu on the formats be explained?

Let us start with the frequency tasks. Recall that in the standard frequency task, the following equation has to be solved:

$$p(H \mid D) = \frac{H \,\&\, D}{H \,\&\, D + \neg H \,\&\, D} = \frac{8}{8 + 95} \tag{1}$$

In the short frequency task, computations are even slightly easier because the sum in the denominator, D (i.e., correct positives and false positives) is already given in the text.

$$p(H \mid D) = \frac{H \,\&\, D}{D} = \frac{8}{103} \tag{2}$$

Thus, only one computational step remains: $H \,\&\, D$ has to be divided by D. The small computational simplification of the short frequency menu, compared to the standard

frequency menu, does not affect performance noticeably, hence the small difference between the performance rates of the two task versions (yet, Gigerenzer & Hoffrage, 1995, interpreted the additional computational step as being consistent with the decrease of performance by four percentage points in the standard frequency task compared to the short frequency task).

For probability and relative frequency formats, on the other hand, the short menu simplifies computations considerably. As seen in Chapter 1, application of Bayes' theorem is needed to solve the standard problem with normalized values such as probabilities or relative frequencies:

$$p(H \mid D) = \frac{p(H)p(D \mid H)}{p(H)p(D \mid H) + p(\neg H)p(D \mid \neg H)} = \frac{(.01)(.80)}{(.01)(.80) + (.99)(.096)} \quad (3)$$

In the short menu, only one operation remains to be performed:

$$p(H \mid D) = \frac{p(H \,\&\, D)}{p(D)} = \frac{.008}{.103} \quad (4)$$

This considerable simplification lead to better performance in the short probability and short relative frequency formats, compared to the respective standard formats.

If one looks at Equations 2 and 4, the actual computations that have to be performed in the short frequency and the short probability tasks look very similar: People have to compute either 8/103 or .008/.103. Thus, the *number* of operations in the short menu is the same for both formats. The difference in performance rates of 22 percentage points between the two formats is attributed to the second aspect of "computational demand", namely, the type of numbers: *"... the short menu is computationally simpler in the frequency than in the probability format, because the frequency format involves calculations with natural numbers and the probability format with fractions"* (Gigerenzer & Hoffrage, 1995, p. 691).

Previous comparisons of short frequency and short probability problems

I found five studies in which short frequency and short probability tasks were compared (Fiedler et al., 2000, Experiment 1; Gigerenzer & Hoffrage, 1995, Study 1; Hoffrage et al., in press, Study 2; Mellers & McGraw, 1999, Studies 1 and 2). The performance rates for the two formats in each study are listed in Table 3.2. As mentioned above, three of the studies

reported a significant difference in the performance rates for the two formats (subsequently called "format effect"), two did not.

Table 3.2
Proportion of correct answers in short frequency and short probability tasks

Study	Short frequency	Short probability
Studies that found a format effect:		
Gigerenzer and Hoffrage (1995)	50%	28%
Mellers and McGraw (1999; rare events)	21%	3%
Hoffrage et al., in press	68%	50%
Studies that did not find a format effect:		
Mellers and McGraw (1999; common events)	16%	16%
Fiedler et al. (2000)	--[a]	--[a]

Note. [a] Fiedler et al. (2000) did not report the proportion of correct answers. They used an absolute deviation measure that will be discussed later in the text.

The first study that reported no format effect in short menu tasks was that of Mellers and McGraw (1999). They expected to find a format effect only in problems with rare events (i.e., problems that include probabilities of .05 or lower), but not in tasks with more common events. They argued that *"most untutored people lack an intuitive feel for the difference between 0.0005 and 0.00005, but this difference is easier to understand when expressed with frequencies as 5 in 10,000 and 5 in 100,000"* (p.419), but that *"when events are more common, the advantage provided by frequencies should be minimized."* (p.420).

Their results supported these predictions. In Study 1, participants received a task with rare events (the mammography problem), and here an advantage of natural frequencies over probabilities was found in the standard and in the short menu (although the absolute level of performance was lower, the relative difference between natural frequency and probability format was similar to that found by Gigerenzer and Hoffrage). In Study 2, participants had to solve a task with more common events (the cab problem); here, no format effect was found and performance was only determined by information menu (short probability 16% vs. standard probability 4% correct; short frequencies 16% vs. standard frequencies 8% correct). Mellers and McGraw (1999) concluded that the commonness of events was an important boundary condition for the facilitating effect of natural frequencies. However, in a reanalysis of their data from 15 different problems, Gigerenzer and Hoffrage (1999) found the predicted interaction of commonness and performance only for the cab problem, and not for the other problems with common events. Hence, the hypothesis that commonness of events is a general boundary condition for the effect of natural frequencies could not be supported, instead the

lack of a format effect reported by Mellers and McGraw (1999) can be explained by specific features of one of the problems they used.

The second study that reported no format effect in short menu tasks was that of Fiedler and colleagues (Fiedler et al., 2000).[5] Adopting the cognitive sampling framework developed by Fiedler (2000), they generally distinguished between two components in statistical inference tasks: an inductive task component of perceiving sample information, and a deductive task component of transforming the input information according to task requirements. If the required transformations are numerous or demanding, biases can occur. In Bayesian inference tasks, the inductive task component is not a problem as long as the tasks provide summary statistics as input information. The problems that occur have to be attributed to the deductive component, that is, to the transformations that have to be performed on the input information to arrive at the solutions. The number of transformations depends on "scale correspondence", that is, on whether the statistical information given has compatible or incompatible reference scales (also referred to as "equal" vs. "unequal scales"). Whether these transformations have to be performed on natural numbers or fractions is not considered relevant as long as the formats do not affect the number of transformations.[6] Accordingly, Fiedler et al. (2000) only found an effect of the factor "scale correspondence" (which was in this case equivalent to information menu; the "unequal scale" tasks corresponded to standard menu tasks, the "equal scale" tasks to short menu tasks), not of the factor statistical format.

Comparison of performance measures and task wordings

I assume that two major differences between the studies of Gigerenzer & Hoffrage (1995) and Fiedler et al. (2000) account for the seemingly contradictory results concerning format effects in short menu tasks.

[5] All references to Fiedler et al. (2000) in the present chapter refer to only the first experiment in their article.
[6] It should be noted that originally, Fiedler et al. had designed the experiment to show that the facilitating effect of natural frequencies reported by Gigerenzer and Hoffrage (1995) was due to a confounded factor. However, the assumption of a confounded factor was based on a misinterpretation: They mistook the standard natural frequency task for a short frequency task (Fiedler et al., 2000, p. 401) and argued that the advantage of natural frequencies in the standard frequency format, compared to the standard probability format (which is termed "unequal scale probability version" by Fiedler et al.) was only due to a difference in the computational demand between the short and the standard version. They therefore added a short probability and a standard natural frequency version ("equal scale probability" and "unequal scale frequency") to disentangle the seemingly confounded factor. However, Gigerenzer and Hoffrage did not confound short and standard menus, but tested a full 2 × 2 factorial design.

The first is the use of different performance measures in the two studies. Gigerenzer and Hoffrage (1995) asked their participants to note their posterior estimates as well as the strategies leading to the estimates. To measure performance, they used a double criterion: A solution was classified correct when (a) the estimate was in a range of ± 1 percentage point of the normative correct posterior value, and (b) the notes of the participant indicated that the estimate actually resulted from a Bayesian strategy (e.g., Equations 1 to 4) rather than guessing. Fiedler et al. (2000) only asked for the posterior estimates. Their performance measure was the deviation of the estimated from the normative correct posterior value. I hypothesize that format effects that are visible in the strategy-based categorical performance measure are not necessarily visible in the estimate-based mean deviation scores. It was unfortunately not possible to get the original data from Fiedler et al. for a reanalysis (although they did not capture strategy data, one could have scored at least the proportion of correct answers according to part (a) of the double criterion for an approximate comparison). I will therefore compare the outcomes of both performance measures in Study 2.

The second major difference between the two studies was the wording of the tasks (Table 3.3). Fiedler et al. wrote that *"to fit this two-factorial frame, the wording of the task had to deviate slightly from that of Gigerenzer and Hoffrage (1995); however, these deviations should not have changed the theoretical predictions"* (Fiedler et al., 2000, p. 403). I would like to test this assumption. In the following, I describe the aspects in which the short menu problems used by Fiedler et al. (2000) differed from the short menu problems used by Gigerenzer and Hoffrage (1995) and discuss the relevance of each change for the prediction of format effects.

Relative frequencies instead of probabilities. In the probability conditions of Fiedler et al., the statistical information given in the task was not represented as single-event probabilities, but as relative frequencies (e.g., "1% of all women have breast cancer", instead of "the probability that a woman has breast cancer is 1%"). Although the labeling is misleading (but not uncommon, e.g. Evans et al., 2000; Macchi, 2000), the use of relative frequencies instead of probabilities per se does *not* affect the predictions by Gigerenzer and Hoffrage (1995). As mentioned earlier, the necessary computations are equivalent for probabilities and relative frequencies, and Gigerenzer and Hoffrage (1995, Study 2) found comparably low performance rates for the two formats in both standard and short menu. Because this was the only study that compared performance with relative frequencies and probabilities in the short

menu[7], I will try to replicate this finding by including both formats in the following two experiments.

Table 3.3
Comparison of the wording of the short menu tasks in the Gigerenzer & Hoffrage (1995) and Fiedler et al. (2000) studies

Condition	Gigerenzer & Hoffrage (1995) (GH short menu)	Fiedler et al. (2000) [a] (FI short menu)
Natural frequency	103 of every 1,000 women at age forty who participate in routine screening will have a positive mammogram. 8 of every 1,000 women at age forty who participate in routine screening have breast cancer *and* a positive mammogram. *Question:* Here is a new representative sample of women in this age group who had a positive mammogram in a routine screening. How many of these women do you expect actually to have breast cancer? ___ out of ___	The study contains data from 1,000 women. 895 women did not have breast cancer and had a negative mammogram; 95 women did not have breast cancer and had a positive mammogram; 8 women had breast cancer and a positive mammogram; and 2 women had breast cancer and a negative mammogram. *Question:* What is the probability of breast cancer, if a woman has a positive mammogram result? The probability is ____%
Probability	The probability that a woman at age forty who participates in routine screening will have a positive mammogram is 10.3%. The probability that a woman at age forty who participates in routine screening will have breast cancer and a positive mammogram is 0.8%. *Question:* A woman in this age group had a positive mammogram in a routine screening. What is the probability that she actually has breast cancer? ____%	The study contains data from 1,000 women. 89.5% of women did not have breast cancer and had a negative mammogram; 9.5% of women did not have breast cancer and had a positive mammogram; 0.8% of women had breast cancer and a positive mammogram; and 0.2% of women had breast cancer and a negative mammogram. *Question:* What is the probability of breast cancer, if a woman has a positive mammogram result? The probability is ____%

Note. [a] The tasks in these conditions were originally labeled "frequency, common reference scale condition" and "probability, common reference scale condition".

Grand total provided in probability / relative frequency task. Fiedler et al. (2000) provided the grand total, that is, the total number of considered cases ("the study contains data from 1,000 women") in the frequency as well as in the probability format (that contained relative frequencies). The grand total is by definition included in the natural frequency tasks, but not necessarily in the probability and relative frequency format. Gigerenzer and Hoffrage (1995) did not provide the grand total or any other absolute frequency in probability or relative frequency tasks. Fiedler et al. intentionally added the grand total to the probability tasks to make them more comparable to the natural frequency tasks, that is, "less misleading"

[7] In the standard menu, the only other study that compared (rather than mixed) the two formats reported a result that is not consistent with the predictions of Gigerenzer and Hoffrage (Macchi, 2000). Here, performance with relative frequencies (37% Bayesian answers) was almost as good as with natural frequencies (40%) and much better than with probabilities (7%).

(2000, p.400, Footnote 2). They argued that the provision of the grand total has a facilitating effect because the grand total makes the common reference scale of the two pieces of information given in short menu tasks explicit. Hoffrage et al. (2002) made a similar prediction on the effect of providing the grand total: *"Providing the total sample (1000 women) serves as a starting point to mimic the procedure of natural sampling, thereby facilitating computational demands considerably. Computing 1% of 1000 women is a simple division that leads automatically to natural frequencies, namely '10 out of 1000 women have breast cancer'. The following statement '80% of the women with breast cancer had a positive mammogram' now directly leads to '8 out of these 10 women have a positive mammogram', etc. The correct answer can now easily be derived – with no danger of confusing conditional probabilities, committing the base-rate fallacy, or struggling with any inversions.'* Thus, the lack of a format effect in the Fiedler et al. study could be attributed to the provision of the grand total in the probability task, a change that makes this task easier than the probability task used by Gigerenzer and Hoffrage (1995).

However, not all invitations to translate normalized information into natural frequencies have been successful in increasing performance. For instance, performance did not increase substantially by asking a frequency question in tasks that provided information in terms of probabilities or relative or normalized frequencies (Girotto & Gonzalez, 2001). Similarly, providing the base rate in terms of frequencies (which also includes the grand total) in standard probability and relative frequency tasks did not increase performance (Macchi, 2000)[8]. Study 3 will address the question whether provision of the grand total in normalized formats is sufficient to increase performance in short probability and relative frequency tasks.

Type of short menu. Fiedler et al. (2000) did not use the same short menu as Gigerenzer and Hoffrage (1995). Whereas Gigerenzer and Hoffrage provided only two pieces of information, namely $H \,\&\, D$ and D (subsequently called "GH short menu"), Fiedler et al. provided all four joint frequencies $H \,\&\, D$, $H \,\&\, \neg D$, $\neg H \,\&\, D$, $\neg H \,\&\, \neg D$ (subsequently called "FI short menu"). In terms of number of computations, the two types are almost equivalent (GH short menu: as in Equation 2; FI short menu: as in Equation 1). However, the FI short menu includes more pieces of information. It is not clear whether and how[9] the amount of

[8] In the standard relative frequency task, Macchi (2000) found 37% Bayesian answers without frequency base rate (she called this version "PP", partitive probability) and 33% Bayesian answers with frequency base rate ("PPbis"). In a standard probability task with frequency base rate, 7% of the participants gave Bayesian answers ("NPP", nonpartitive probability); this performance rate is not higher than the performance rates found in other studies for standard probability versions without frequency base rates (no such version was included in Macchi, 2000).

[9] On the one hand, fewer information units imply lower attentional demand which could facilitate solving the task. On the other hand, there is evidence that the provision of the complementary values of relevant statistical

information displayed in a Bayesian inference task could affect performance (that is, independent of the number of computations). To control for this potential source of variation, Study 2 will include both types of short menus.

Question format. Fiedler et al. used the same statistical format in the question for every task, that is, each question asked for a single-event probability. Therefore, also participants who received information in terms of natural frequencies were asked to make a single-event point estimate in the end. Gigerenzer and Hoffrage (1995) did not mix formats in the natural frequency and probability problems (they did, however, use a probability question in the relative frequency problems, see Table 1.1). How do mixed formats affect performance? For a slightly different version of Bayesian inference problems[10], Cosmides and Tooby (1996) showed that performance was impaired when the information was represented in natural frequencies while the question asked for a probability. And as already mentioned above, in one study a frequency question in a relative frequency problem lead to even worse performance compared to relative frequency problems without frequency question (Girotto & Gonzalez, 2001). To assess the effect of genuine natural frequency, relative frequency and probability formats on performance in short menu problems, the formats of statistical information and question were consistent in all inference problems used in Study 2 and 3.

To conclude, the comparison of performance measures and task wordings in the studies of Gigerenzer and Hoffrage (1995) and Fiedler et al. (2000) lead to a number of open questions and hypotheses that could account for the seemingly inconsistent results. Four of the hypotheses will be followed up on Study 2 and 3.

Study 2

The goals of Study 2 are the following: First, it tests whether performance in short frequency tasks is higher than in short probability and short relative frequency tasks. Finding a format effect would support the notion of Gigerenzer and Hoffrage (1995) that natural frequencies as cardinal numbers are easier to process than percentages and probabilities as fractions, even if the number of computations is the same (and would that replicate the findings of Gigerenzer and Hoffrage, 1995; Gigerenzer et al., in press). Finding no format effect would support the

information can, although redundant, improve judgments (e.g., people are less susceptible to framing manipulations when the two complementary outcomes of an option are rendered explicit; Kühberger, 1995; another example is the reduction of acquiescence with two-sided questions, see Bishop, Oldendick, & Tuchfarber, 1982).

[10] The sensitivity was set to 100%, thus only the false-alarm rate had to be taken into account for the calculation of the positive predictive value.

notion of Fiedler et al. (2000) that only the number of necessary computations determines performance in Bayesian inference tasks. I will compare short frequency with short probability *and* short relative frequency tasks to replicate the finding of Gigerenzer and Hoffrage (1995) that the latter two formats yield equivalent performance rates. Second, it tests the hypothesis that format effects that are visible in the strategy-based, categorical performance measure are not necessarily visible in the estimate-based mean deviation scores. Evidence for this hypothesis would suggest that the conclusions of Gigerenzer and Hoffrage and Fiedler et al. concerning format effects in Bayesian inference tasks are not contradicting, but just not comparable. Third, it explores whether the type of short menu (FI or GH short menu) has any effect on performance.

Method

Design

Participants were randomly assigned to one of three conditions. In all three conditions, they worked on four short menu problems of the same format, that is, either on short frequency, short probability, or short relative frequency tasks. Two of the four tasks in each condition had a GH short menu, the other two a FI short menu. Thus, a 3 (statistical format) × 2 (type of short menu) mixed design was used. The order of the menus and the order of the four problems (mammography problem, lung damage problem, colorectal cancer problem, Down's syndrome problem) were partially randomized. Following the 3 × 2 design, six versions of each problem were constructed. Table 3.4 shows the six versions for the mammography problem, the other problems can be found in the Appendix.

Procedure

Participants worked on average 40 minutes on the four problems. There was no time limit. They were asked not only to write down their solutions for the problems, but also to document the strategies they used to arrive at the solutions. After working through the four problems, participants received a short questionnaire in which they were asked to what extent they had previously been familiar with Bayes' theorem. Upon completion, they were paid a flat fee of 10 Euro.

Participants

The participants were 60 students, 44 women and 16 men, from various disciplines (predominantly psychology) at the Free University of Berlin. Sixteen of the participants said that they had heard of Bayes' theorem before. However, 15 of the 16 said that they could not remember it. The one student who said he could remember it did not profit from this knowledge, because he could not solve any of the four tasks correctly.

Table 3.4
The 6 versions of the mammography problem used in Study 2

Format	GH short menu	FI short menu
Natural frequencies	103 out of every 1,000 women receive a positive mammogram. 8 out of every 1,000 women have breast cancer *and* receive a positive mammogram.	8 out of every 1,000 women have breast cancer and receive a positive mammogram. 2 out of every 1,000 women have breast cancer and receive a negative mammogram. 895 out of every 1,000 women do not have breast cancer and receive a negative mammogram. 95 out of every 1,000 women do not have breast cancer and receive a positive mammogram.
	Question: Imagine a new representative sample of women (without symptoms, same age group) who all received a positive mammogram in routine screening. How many of these women actually have breast cancer? ___ out of ___	
Relative frequencies	10.3% of women receive a positive mammogram. 0.8% of women have breast cancer and receive a positive mammogram.	0.8% of women have breast cancer and receive a positive mammogram. 0.2% of women have breast cancer and receive a negative mammogram. 89.5% of women do not have breast cancer and receive a negative mammogram. 9.5% of women do not have breast cancer and receive a positive mammogram.
	Question: Imagine a new representative sample of women (without symptoms, same age group) who all received a positive mammogram in routine screening. What percentage of these women actually has breast cancer? ____%	
Probabilities	The probability that one of these women receives a positive mammogram is 10.3%. The probability that one of these women has breast cancer and receives a positive mammogram is 0.8%.	The probability that one of these women has breast cancer and receives a positive mammogram is 0.8%. The probability that one of these women has breast cancer and receives a negative mammogram is 0.2%. The probability that one of these women does not have breast cancer and receives a negative mammogram is 89.5%. The probability that one of these women does not have breast cancer and receives a positive mammogram is 9.5%.
	Question: A woman in this age group had a positive mammogram in routine screening. What is the probability that she actually has breast cancer? ____%	

Note. The short introductory text that preceded the problems can be found in the Appendix.

Results

The first question was whether the statistical format had an effect on performance in the short menu. The answer is yes. Natural frequencies elicited a substantially higher proportion of correct Bayesian inferences than relative frequencies and probabilities (see Table 3.5). Across both short menu types, natural frequencies elicited 60% correct inferences, compared to 29% in the relative frequency and 23% in the probability format ($\chi^2(1, 160) =$

15.83, $p < .01$, $\phi = .31$, and $\chi^2(1, 160) = 23.21$, $p < .01$, $\phi = .38$, respectively). Thus, the effect of statistical format was of "medium" size according to the classification of Cohen (1988; see Chapter 2). Performance in the short relative frequency tasks was, with 29% correct answers, only slightly better than in the short probability tasks with 23%, $\chi^2(1, 160) = 0.82$, $p > .05$, $\phi = .07$.

Table 3.5
Proportion of correct Bayesian inferences in Study 2

	GH short menu	FI short menu	*Total*
Natural frequencies	60% (24)	60% (24)	*60% (48)*
Relative frequencies	30% (12)	28% (11)	*29% (23)*
Probabilities	23% (9)	23% (9)	*23% (18)*
Total	*38% (45)*	*38% (45)*	

Note. The percentages refer to N = 40 tasks per cell (except for the totals). Values in parentheses are absolute numbers of correct inferences per cell.

Table 3.5 also shows that there was almost no difference between the two short menu types. In the GH short menu tasks, 45 of 120 tasks were solved correctly, compared to 44 of 120 FI short menu tasks.

Another question was whether format effects that are visible in the above categorical performance measure are also visible in the estimate-based mean deviation scores. Table 3.6 shows the results for the mean deviation scores used by Fiedler et al., pooled for the two types of menus. The highly significant format effect reported above is not visible in these estimates, and ANOVAs for the four problems did not detect any significant differences between the statistical formats. One could question whether the mean deviation scores reported by Fiedler et al. is at all suited to detect differences here, because usually over- *and* underestimations of the correct posterior probability can be found in the participants' answers that cancel each other out when summed up. I therefore also analyzed the mean *absolute* deviations. However, the result was the same as before, no significant differences between the statistical formats based on deviation scores (Table 3.7).

Table 3.6
Mean deviations of participants' posterior probability estimates from the correct value in Study 2

Problem content	Natural frequencies	Relative frequencies	Probabilities	*F (df)*
Mammography	- 1.44%	2.47%	1.39%	0.28 (2)
Colorectal cancer	- 0.83%	7.21%	4.52%	0.91 (2)
Lung disease	1.54%	1.64%	- 1.09%	1.10 (2)
Down's syndrome	3.47%	- 0.36%	8.15%	1.07 (2)

Note. The mean deviations refer to N= 20 estimates per cell.

Table 3.7
Mean *absolute* deviations of participants' posterior probability estimates from the correct value in Study 2

Problem content	Natural frequencies	Relative frequencies	Probabilities	*F (df)*
Mammography	2.01%	8.12%	7.88%	0.99 (2)
Colorectal cancer	0.89%	12.31%	9.30%	2.24 (2)
Lung disease	1.97%	2.92%	1.09%	0.41 (2)
Down's syndrome	3.96%	0.90%	9.69%	1.22 (2)

Note. The mean deviations refer to N= 20 estimates per cell.

Summary

Study 2 showed that significantly more people solved the short menu tasks correctly when the problems were stated in natural frequencies than in percentages or probabilities. However, this format effect was not visible in the second performance measure, mean deviations and mean absolute deviations from the normative correct value. Performance in the short relative frequency and short probability format was comparably low. Performance rates did also not differ between the two tested short menu types.

Study 3

Study 3 assessed the impact of the grand total in short probability and short relative frequency tasks. According to Fiedler et al (2000) and Hoffrage et al. (2002), performance in these two formats should be better when the grand total is given than when it is not given. If the data show such an effect of the grand total, this would suggest that the contradicting results in the studies of Fiedler et al. (2000) and Gigerenzer and Hoffrage (1995) could also be caused by the differing use of the grand total in probability tasks.

None of the explanations of the effect of the grand total mentioned above suggested a differential effect of the grand total on short probability and short relative frequency tasks. However, it is not implausible to assume that the grand total could reduce uncertainty in the two formats to differing degrees. As the quote by Hoffrage et al. (see above) indicates, relating relative frequencies to the grand total leads to a new absolute number. For instance, if 1% of all women in the considered sample have breast cancer and 1,000 women are in the sample, then 10 women in this sample have breast cancer. In the probability format, however, the number of women with breast cancer in the sample is not clear even with a grand total: If the probability for a woman in the sample to have breast cancer is 1% and the sample consists of 1,000 women, the expected number of breast cancer cases is 10, but the actual number

could be higher or lower. There is still more uncertainty in this case. Therefore, I included both short probability and short relative frequency tasks in Study 3 to check for differential effects of the grand total.

Method

Design

Participants were randomly assigned to one of two conditions. In both conditions, they worked on two short menu problems that both either did or did not provide the grand total (the factor provision of grand total was realized as a between-participants factor to avoid transfer effects). One of the two tasks in each condition had a relative frequency format, the other had a probability format. Thus, in contrast to Study 2, the factor statistical format was realized here as a within-participant factor. But as Study 2 showed that performance in the two conditions was comparable, no unwanted transfer effects have to be expected. Hence, a 2 (statistical format) × 2 (provision of grand total) mixed design was used.

The order of the formats and the order of the two problems (mammography problem, Down's syndrome problem) were partially randomized. Following the 2 × 2 design, four versions of each problem were constructed. Because there was no effect of short menu type in Study 2, only the GH short menu was used. The tasks without grand total were the same as in Study 2. In the tasks with grand total, the statistical information was preceded by the sentence "A study that examined this question contained data from 1,000 women [in the mammography problem] / 10,000 pregnant women [in the Down's syndrome problem]".

Procedure

The procedure was the same as in Study 2. Participants worked on average 20 minutes on the two problems and were paid a flat fee of 5 Euro upon completion.

Participants

The participants were 50 students, 28 women and 22 men, from various disciplines (predominantly psychology) from the Free University of Berlin. Fifteen of the participants said that they had heard of Bayes' theorem before. However, 14 of the 15 said that they could not remember the theorem. The one student who could remember it solved both tasks correctly.

Results

Did the provision of the grand total increase performance in the two short menu tasks? The answer is no. Across the two formats, participants solved 15 of 50 tasks without grand total correctly, and 12 of 50 tasks with grand total (Table 3.8). There were also no notable

differences within the two formats. Within the relative frequency format, the number of correct solutions was the same with or without grand total. Within short probability format, only few more people solved the task correctly when the grand total was *not* given (8 vs. 5 correct answers). The absolute level of performance for the short percentage and short probability tasks with grand total (28% and 20% correct, respectively) did not reach that for the short frequency menu in Study 2 (60% correct).

Table 3.8
Proportion of correct Bayesian inferences in Study 3

Format	Without grand total	With grand total	*Total*
Relative frequencies	28% (7)	28% (7)	*28% (14)*
Probabilities	32% (8)	20% (5)	*26% (13)*
Total	*30% (15)*	*24% (12)*	

Note. The proportion of correct Bayesian inferences refers to N = 25 tasks per cell (except for the totals). Values in parentheses are absolute numbers of correct inferences per cell.

When the grand total was given, only two more tasks were solved correctly in the relative frequency format than in the probability format (7 vs. 5 correct answers). Thus, there was no differential effect of the grand total on the two formats.

To check the prediction of Hoffrage et al. (2002) that the provision of the grand total prompts participants to translate percentages and probabilities into natural frequencies, I categorized the protocols of the solution process made by the participants into three groups. (a) *No translation* took place when the participant operated exclusively with the numbers given in the task, that is, percentages and probabilities. If participants who received tasks with the grand total simply reproduced this number in the notes, but did not use it as a basis for further operations, this was also scored as No translation. (b) *Translation* took place when the participants actively translated all or part of the information given into absolute frequencies and used this frequency information to answer the task. (c) *Back-and-forth* took place when the participants switched between frequency and percentage representations. For instance, in some answers in this category the participant had started to translate the information given into frequencies, but then for some reason went back to the given percentages and summed them up to answer the question.

As Table 3.9 shows, the proportion of translations in the tasks with grand total only was slightly higher than in the tasks without grand total (16 vs. 10 out of 50 tasks). There was almost no difference between the relative frequency tasks with and without grand total (9 vs. 7 translations), $\chi^2(1, 50) = .37$, $p > .05$, $\phi = .09$, and only a small difference between the two probability tasks (7 vs. 3 translations), $\chi^2(1, 50) = 2.00$, $p > .05$, $\phi = .20$. The most frequent

strategy in all four conditions was "No translation" (53% of all tasks). The "Back-and-forth" pattern could be found in 21% of all tasks.

Table 3.9
Strategy use in the four conditions of Study 3 and proportion of correct answers per strategy

Condition		Translations	No translations	Back-and-forth
Without grand total	- Relative frequencies	28% (7)	52% (13)	20% (5)
	- Probabilities	12% (3)	72% (18)	16% (4)
With grand total	- Relative frequencies	36% (9)	44% (11)	20% (5)
	- Probabilities	28% (7)	44% (11)	28% (7)
Total		*26*	*53*	*21*
Proportion correct [a]		54% (14)	15% (8)	24% (5)

Note. N = 25 in each of the four conditions. Values in parentheses are absolute numbers of strategy use / correct inferences per cell. [a] The percentages of correct answers refer to the column above, the total number of participants who used the strategy.

The process of translating percentages and probabilities into natural frequencies while solving the tasks was closely linked to performance. The bottom row of Table 3.9 shows that the proportion of correct answers was highest when participants translated the percentage and probability information into frequencies (14 of 26 tasks). The differences in the number of correct answers of the translation group compared to the other two groups were of medium effect size (Translation vs. No translation: $\chi^2(1, 79) = 13.04$, $p < .01$, $\phi = .41$; Translation vs. Back-and-forth: $\chi^2(1, 47) = 4.35$, $p < .05$, $\phi = .30$).

Summary

Study 3 showed that the mere addition of the grand total to short relative frequency and short probability tasks did not increase performance. Even with grand total, performance in both formats remained about 30 percentage points lower than in the short natural frequency format. The provision of the grand total led only to a small increase in translations into natural frequencies for the probability tasks. Whenever a translation into natural frequencies was performed, however, participants were more likely to arrive at the correct solution than without translation.

Summary and discussion

The main question of the present chapter was whether there are format effects in short menu tasks. In short menu tasks, the number of computations is the same in the natural frequency, relative frequency and probability format. Thus, if natural frequencies still had an advantage

over the other formats in short menu tasks, this would support the notion that the facilitating effect of natural frequencies is also based on the fact that they are cardinal numbers rather than fractions. This feature is relevant for applications of the tool of natural frequencies outside the text problem paradigm.

Therefore the goal of this chapter was to clarify why there were contradicting results concerning format effects in short menu tasks, and whether these findings implied boundary conditions for the effectiveness of natural frequencies in the risk communication context. I focused on two studies that did not find a format effect. The lack of a format effect in the first study (Mellers & McGraw, 1999) could be attributed to features of the specific text problem used (Gigerenzer & Hoffrage, 1999), so that there is no evidence for a general boundary condition for the risk communication context.

For the second study (Fiedler et al., 2000), not one but several factors could potentially explain the differing results. Study 2 and 3 ruled out that the assumptions that two deviations in the task wording were responsible for the lack of format effect (type of short menu and provision of grand total).

But maybe the results of the Fiedler et al. and the Gigerenzer and Hoffrage study were not contradictory in the first place. The reason is that the studies used different measures to assess performance. As Study 2 showed, the fact that significantly more Bayesian strategies were used in the short frequency tasks (better performance according to Gigerenzer and Hoffrage) did not lead to a significantly lower mean deviation of the estimates from the correct value in this condition (which would have meant better performance according to Fiedler et al.). One reason for this observation could be that statistical format did not only influence the proportion of correct strategies, but also the distribution of incorrect strategies (namely that incorrect strategies yielding large deviations from the correct value were more frequent in the natural frequency tasks than in the other tasks). In any case the result underlines how important it is that different studies on the same subject use the same performance measures so that results can be compared across studies. In fact, most studies on format effects in Bayesian inference tasks used categorical performance measures, although not all of them analyzed participant's strategies in addition to their estimates (e.g., Girotto & Gonzalez, 2001; Macchi, 2000). This is not to say that deviation-based or other performance measures were less interesting. Obviously, the choice of dependent measures has to be based on the goals of the study. But when one of these goals is to compare the own results to those of previous studies, then at least one of the measures should be comparable. To conclude, due to different performance measures, it is not possible to determine whether the results of

Fiedler et al. (2000) did at all contradict those of Gigerenzer and Hoffrage (1995). Thus, also this study did not provide evidence for a boundary condition for the facilitating effect of natural frequencies.

However, there is one potentially relevant factor that I addressed in the analysis of the task wording, but did not test in the experiments. It could still be that Fiedler et al. did indeed not find a format effect because their frequency tasks were more difficult than those of Gigerenzer and Hoffrage due to the use of single-event probability questions. If the question format would have played a role (and as mentioned above, there is some evidence for this hypothesis, see Macchi, 2000), then what would this mean for the application of natural frequencies in medical risk communication? The implication is that risk communicators and designers of public health information materials should be cautious to use more than one format in the same conversation or text. This is consistent with the more general fact that people get easily confused when they have to deal with differently scaled numerical information (e.g., Fiedler, 2000; Gigerenzer, 2002; Krämer, 1991). But in the specific context of diagnostic inference problems, there is yet too little knowledge about when mixing formats hinders statistical thinking and when it does not (but see Cosmides & Tooby, 1996; Girotto & Gonzalez, 2001). To answer this question, more systematic reviews of all studies that – intentionally or unintentionally – used mixed formats are needed.

Some of the tasks in Study 3 also "mixed" formats by supplementing relative frequency or probability information with the grand total, the absolute number of all considered cases. The results suggest that the mere provision of the grand total is not sufficient to increase performance in short relative frequency and short probability tasks. The strategy to translate normalized information into natural frequencies by using the grand total as a starting point was very successful, but without further instruction (as, for instance, in Chapter 2), only few people used this strategy.

Thus, without the possibility to give further instructions, the most promising approach to foster understanding of medical risk communication is to represent the statistical information in terms of natural frequencies. The results of this chapter support the notion of Gigerenzer and Hoffrage (1995) that the beneficial effect of natural frequencies is, among others, also due to the fact that they are cardinal numbers and as such easier to deal with than fractions. Moreover, they always specify and thus disambiguate the reference class. The following chapters will explore how these features can be used to facilitate comprehension in a specific case of medical risk communication, namely information about mammography screening.