

# Probabilistic Mental Models: A Brunswikian Theory of Confidence<sup>1</sup>

**Gerd Gigerenzer**

Center for Advanced Study in the Behavioral Sciences, Stanford, California

**Ulrich Hoffrage and Heinz Kleinbölting**

University of Salzburg, Salzburg, Austria

Research on people's confidence in their general knowledge has to date produced two fairly stable effects, many inconsistent results, and no comprehensive theory. We propose such a comprehensive framework, the theory of probabilistic mental models (PMM theory). The theory (a) explains both the overconfidence effect (mean confidence is higher than percentage of answers correct) and the hard-easy effect (overconfidence increases with item difficulty) reported in the literature and (b) predicts conditions under which both effects appear, disappear, or invert. In addition, (c) it predicts a new phenomenon, the confidence-frequency effect, a systematic difference between a judgment of confidence in a single event (i.e., that any given answer is correct) and a judgment of the frequency of correct answers in the long run. Two experiments are reported that support PMM theory by confirming these predictions, and several apparent anomalies reported in the literature are explained and integrated into the present framework.

Do people think they know more than they really do? In the last 15 years, cognitive psychologists have amassed a large and apparently damning body of experimental evidence on overconfidence in knowledge, evidence that is in turn part of an even larger and more damning literature on so-called cognitive biases. The cognitive bias research claims that people are naturally prone to making mistakes in reasoning and memory, including the mistake of overestimating their knowledge. In this article, we propose a new theoretical model for confidence in knowledge based on the more charitable assumption that people are good judges of the reliability of their knowledge, provided that the knowledge is representatively sampled from a specified reference class. We claim that this model both predicts new experimental results (that we have tested) and explains a wide range of extant experimental findings on confidence, including some perplexing inconsistencies.

Moreover, it is the first theoretical framework to integrate the two most striking and stable effects that have emerged from confidence studies—the overconfidence effect and the hard-easy effect—and to specify the conditions under which these effects can be made to appear, disappear, and even invert. In most recent studies (including our own, reported herein), subjects are asked to choose between two alternatives for each of a series of general-knowledge questions. Here is a typical example: “Which city has more inhabitants? (a) Hyderabad or (b) Islamabad.” Subjects

---

<sup>1</sup> This article was written while Gerd Gigerenzer was a fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, California. We are grateful for financial support provided by the Spencer Foundation and the Deutsche Forschungsgemeinschaft (DFG 170/2-1).

We thank Leda Cosmides, Lorraine Daston, Baruch Fischhoff, Jennifer Freyd, Kenneth Hammond, Wolfgang Hell, Sarah Lichtenstein, Kathleen Much, John Tooby, Amos Tversky, and an anonymous reviewer for helpful comments on earlier versions of this article.

choose what they believe to be the correct answer and then are directed to specify their degree of confidence (usually on a 50%–100% scale) that their answer is indeed correct. After the subjects answer many questions of this sort, the responses are sorted by confidence level, and the relative frequencies of correct answers in each confidence category are calculated. The *overconfidence effect* occurs when the confidence judgments are larger than the relative frequencies of the correct answers; the *hard-easy effect* occurs when the degree of overconfidence increases with the difficulty of the questions, where the difficulty is measured by the percentage of correct answers.

Both effects seem to be stable. Fischhoff (1982) reviewed the attempts to eliminate overconfidence by numerous “debiasing methods,” such as giving rewards, clarifying instructions, warning subjects in advance about the problem, and using better response modes—all to no avail. He concluded that these manipulations “have so far proven relatively ineffective” and that overconfidence was “moderately robust” (p. 440) von Winterfeldt and Edwards (1986, p. 539) agreed that “overconfidence is a reliable, reproducible finding.” Yet these robust phenomena still await a theory. In particular, we lack a comprehensive theoretical framework that explains both phenomena, as well as the various exceptions reported in the literature, and integrates the several local explanatory attempts already advanced. That is the aim of this article. It consists of four parts: (a) an exposition of the proposed theory of probabilistic mental models (PMM theory), including predictions of new experimental findings based on the theory; (b) a report of our experimental tests confirming these predictions; (c) an explanation of apparent anomalies in previous experimental results, by means of PMMs; and (d) a concluding discussion.

## PMM Theory

This theory deals with spontaneous confidence—that is, with an immediate reaction, not the product of long reflection. Figure 1 shows a flow chart of the processes that generate confidence judgments in two-alternative general-knowledge tasks.<sup>2</sup> There are two strategies. When presented with a two-alternative confidence task, the subject first attempts to construct what we call a *local mental model* (local MM) of the task. This is a solution by memory and elementary logical operations. If this fails, a PMM is constructed that goes beyond the structure of the task in using probabilistic information from a natural environment.

For convenience, we illustrate the theory using a problem from the following experiments: “Which city has more inhabitants? (a) Heidelberg or (b) Bonn.” As explained earlier, the subjects’ task is to choose a or b and to give a numerical judgment of their confidence (that the answer chosen is correct).

### *Local MM*

We assume that the mind first attempts a direct solution that could generate certain knowledge by constructing a local MM. For instance, a subject may recall from memory that Heidelberg has a population between 100,000 and 200,000, whereas Bonn has more than 290,000 inhabitants. This is already sufficient for the answer “Bonn” and a confidence judgment of 100%. In general,

---

<sup>2</sup> For convenience, the theory is presented here in its complete form, although parts of it were developed after Experiment 1 was performed. All those parts were subjected to an independent test in Experiment 2.

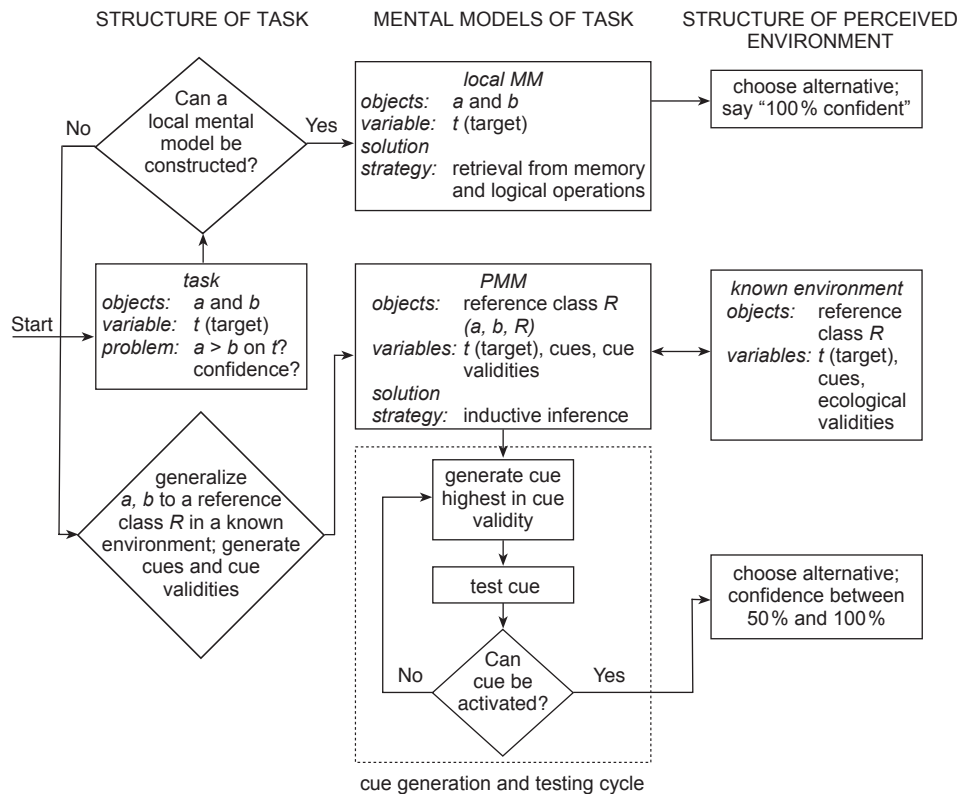


Figure 1. Cognitive processes in solving a two-alternative general-knowledge task (MM = mental model; PMM = probabilistic mental model).

a local MM can be successfully constructed if (a) precise figures can be retrieved from memory for both alternatives, (b) intervals that do not overlap can be retrieved, or (c) elementary logical operations, such as the method of exclusion, can compensate for missing knowledge. Figure 2 illustrates a successful local MM for the previous example. Now consider a task where the target variable is not quantitative (such as the number of inhabitants) but is qualitative: “If you see the nationality letter *P* on a car, is it from Poland or Portugal?” Here, either direct memory about the correct answer or the method of exclusion is sufficient to construct a local MM. The latter is illustrated by a subject reasoning “Since I know that Poland has PL it must be Portugal” (Allwood & Montgomery, 1987, p. 370).

The structure of the task must be examined to define more generally what is referred to as a local MM. The task consists of two objects, *a* and *b* (alternatives), and a target variable *t*. First, a local MM of this task is local; that is, only the two alternatives are taken into account, and no reference class of objects is constructed (see the following discussion). Second, it is direct; that is, it contains only the target variable (e.g., number of inhabitants), and no probability cues are used. Third, no inferences besides elementary operations of deductive logic (such as exclusion) occur. Finally, if the search is successful, the confidence in the knowledge produced is evaluated as certain. In these respects, our concept of a local MM is similar to what Johnson-Laird (1983, pp. 134–142) called a “mental model” in syllogistic inference.

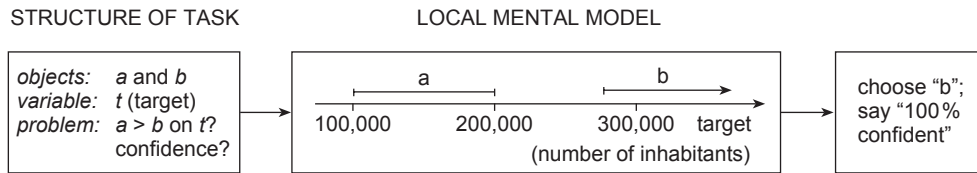


Figure 2. Local mental model of a two-alternative general-knowledge task.

A local MM simply matches the structure of the task; there is no use of the probability structure of an environment and, consequently, no frame for inductive inference as in a PMM. Because memory can fail, the “certain” knowledge produced can sometimes be incorrect. These failures contribute to the amount of overconfidence to be found in 100%-confident judgments.

### PMM

Local MMs are of limited success in general-knowledge tasks<sup>3</sup> and in most natural environments, although they seem to be sufficient for solving some syllogisms and other problems of deductive logic (see Johnson-Laird, 1983). If no local MM can be activated, it is assumed that a PMM is constructed next. A PMM solves the task by inductive inference, and it does so by putting the specific task into a larger context. A PMM connects the specific structure of the task with a probability structure of a corresponding natural environment (stored in long-term memory). In our example, a natural environment could be the class of all cities in Germany with a set of variables defined on this class, such as the number of inhabitants. This task selects the number of inhabitants as the target and the variables that covary with this target as the cues.

A PMM is different from a local MM in several respects. First, it contains a *reference class* of objects that includes the objects *a* and *b*. Second, it uses a network of variables in addition to the target variable for indirect inference. Thus, it is neither local nor direct. These two features also change the third and fourth aspects of a local MM. Probabilistic inference is part of the cognitive process, and uncertainty is part of the outcome.

### Reference Class

We use Brunswik’s (1943, p. 257) term *reference class* to define the class of objects or events that a PMM contains. In our example, the reference class “all cities in Germany” may be generated. To generate a reference class means to generate a set of objects known from a person’s natural environment that contains objects *a* and *b*.

The reference class determines which cues can function as probability cues for the target variable and what their cue validities are. For instance, a valid cue in the reference class “all cities in Germany” would be the soccer-team cue; that is, whether a city’s soccer team plays in the German soccer Bundesliga, in which the 18 best teams compete. Cities with more inhabitants are more

<sup>3</sup> Allwood and Montgomery (1987, pp. 369–370) estimated from verbal protocols that about 19% of their general-knowledge questions were solved by “full recognition,” which seems to be equivalent to memory and elementary logical operations only.

likely to have a team in the Bundesliga. The soccer-team cue would not help in the Hyderabad-Islamabad task, which must be solved by a PMM containing a different reference class with different cues and cue validities.

### *Probability Cues*

A PMM for a given task contains a reference class, a target variable, probability cues, and cue validities. A variable is a probability cue  $C_i$  (for a target variable in a reference class  $R$ ) if the probability  $p(a)$  of  $a$  being correct is different from the conditional probability of  $a$  being correct, given that the values of  $a$  and  $b$  differ on  $C_i$ . If the cue is a binary variable such as the soccer-team cue, this condition can be stated as follows:

$$p(a) \neq p(a \mid aC_i b; R),$$

where  $aC_i b$  signifies the relation of  $a$  and  $b$  on the cue  $C_i$  (e.g.,  $a$  has a soccer team in the Bundesliga, but  $b$  does not) and  $p(a \mid aC_i b; R)$  is the cue validity of  $C_i$  in  $R$ .

Thus, cue validities are thought of as conditional probabilities, following Rosch (1978) rather than Brunswik (1955), who defined his “cue utilizations” as Pearson correlations. Conditional probabilities need not be symmetric as correlations are. This allows the cue to be a better predictor for the target than the target is for the cue, or vice versa. Cue validity is a concept in the PMM, whereas the corresponding concept in the environment is *ecological validity* (Brunswik, 1955), which is the true relative frequency of any city having more inhabitants than any other one in  $R$  if  $aC_i b$ . For example, consider the reference class *all cities in Germany with more than 100,000 inhabitants*. The ecological validity of the soccer-team cue here is .91 (calculated for 1988/1989 for what then was West Germany). That is, if one checked all pairs in which one city  $a$  has a team in the Bundesliga but the other city  $b$  does not, one would find that in 91% of these cases city  $a$  has more inhabitants.

### *Vicarious Functioning*

Probability cues are generated, tested, and if possible, activated. We assume that the order in which cues are generated is not random; in particular, we assume that the order reflects the hierarchy of cue validities. For the reference class *all cities in Germany*, the following cues are examples that can be generated: (a) the soccer-team cue; (b) whether one city is a state capital and the other is not (state capital cue); (c) whether one city is located in the Ruhrgebiet, the industrial center of Germany, and the other in largely rural Bavaria (industrial cue); (d) whether the letter code that identifies a city on a license plate is shorter for one city than for the other (large cities are usually abbreviated by only one letter, smaller cities by two or three; license plate cue); and (e) whether one has heard of one city and not of the other (familiarity cue). Consider now the Heidelberg-Bonn problem again. The first probability cue is generated and tested to see whether it can be activated for that problem. Because neither of the two cities has a team in the Bundesliga, the first cue does not work.

In general, with a binary cue and the possibility that the subject has no knowledge, there are nine possibilities (see Figure 3). In only two of these can a cue be activated. In all other cases, the cue is useless (although one could further distinguish between the four known-unknown cases

		City a Soccer team in Bundesliga?		
		yes	no	unknown
City b Soccer team in Bundesliga?	yes			
	no			
	unknown			

*Figure 3.* Two conditions in which a cue can be activated.

and the three remaining cases). If a cue cannot be activated, then a further cue is generated and tested. In the Heidelberg-Bonn task, none of the five cues cited earlier can in fact be activated. Finally, one cue may be generated that can be activated, such as whether one city is the capital of the country and the other is not (capital cue). This cue has a small probability of being activated—a small activation rate in  $R$  (because it applies only to pairs that include Bonn)—and it does not have a particularly high cue validity in  $R$  because it is well-known that Bonn is not exactly London or Paris.

The Heidelberg-Bonn problem illustrates that probability cues may have small activation rates in  $R$ , and as a consequence, several cues may have to be generated and tested before one is found that can be activated. The capital cue that can be activated for the Heidelberg-Bonn comparison may fail for the next problem, for instance a Heidelberg-Göttingen comparison. Cues can substitute for one another from problem to problem, a process that Brunswik (1955) called “vicarious functioning.”

### *End of Cue Generation and Testing Cycle*

If (a) the number of problems is large or other kinds of time pressure apply and (b) the activation rate of cues is rather small, then one can assume that the cue generation and testing cycle ends after the first cue that can be activated has been found. Both conditions seem to be typical for general-knowledge questions. For instance, even when subjects were explicitly instructed to produce all possible reasons for and against each alternative, they generated only about three on the average and four at most (Koriat, Lichtenstein, & Fischhoff, 1980). If no cue can be activated, we assume that choice is made randomly, and “confidence 50%” is chosen.

*Choice of Answer and Confidence Judgment*

Choice of answer and confidence judgment are determined by the cue validity. Choice follows the rule:

$$\text{choose } a \text{ if } p(a | aC_i b; R) > p(b | aC_i b; R).$$

If  $a$  is chosen, the confidence that  $a$  is correct is given by the cue validity:

$$p(a | aC_i b; R).$$

Note that the assumption that confidence equals cue validity is not arbitrary; it is both rational and simple in the sense that good calibration is to be expected if cue validities correspond to ecological validities. This holds true even if only one cue is activated.

Thus, choice and confidence are inferred from the same activated cue. Both are expressions of the same conditional probability. Therefore, they need not be generated in the temporal sequence choice followed by confidence. The latter is, of course, typical for actual judgments and often enforced by the instructions in confidence studies.

*Confidence in the Long Run and Confidence in Single Events*

Until now, only confidence in single events—such as the answer “Bonn” is correct—has been discussed. Confidence in one’s knowledge can also be expressed with respect to sequences of answers or events, such as “How many of the last 50 questions do you think you answered correctly?” This distinction is parallel to that between probabilities of single events and relative frequencies in the long run—a distinction that is fundamental to all discussions on the meaning of probability (see Gigerenzer et al., 1989). Probabilities of single events (confidences) and relative frequencies are not the same for many schools of probability, and we argue that they are not evaluated by the same cognitive processes either.

Consider judgments of frequency. General-knowledge tasks that involve a judgment of the frequency of correct answers (frequency tasks) can rarely be answered by constructing a local MM. The structure of the task contains one sequence of  $N$  questions and answers, and the number of correct answers is the target variable. Only limiting cases, such as small  $N$  (i.e., if only a few questions are asked) combined with the belief that all answers were correct, may allow one to solve this task by a local MM. Again, to construct a local MM of the task means that the mental model consists of only the local sequence of total  $N$  answers (no reference class), and because one attempts to solve the task by direct access to memory about the target variable, no network of probability cues is constructed.

Similarly, a PMM of a frequency task is different from a PMM of a confidence task. A confidence task about city size in Germany has “cities in Germany” as a reference class; however, a task that involves judgments of frequencies of correct answers in a series of  $N$  questions about city size has a different reference class: Its reference class will contain series of similar questions in similar testing situations. Because the target variable also differs (number of correct answers instead of number of inhabitants), the PMM of a frequency task will also contain different cues and cue validities. For instance, base rates of performance in earlier general knowledge or similar testing situations could serve as a probability cue for the target variable. Again, our basic assumption is that a PMM connects the structure of the task with a known structure of the subject’s environment.

Table 1  
*Probabilistic Mental Models for Confidence Task Versus Frequency Task:  
 Differences Between Target Variables, Reference Classes, and Probability Cues*

PMM	Confidence task	Frequency task
Target variable	Number of inhabitants	Number of correct answers
Reference class	Cities in Germany	Sets of general-knowledge questions in similar testing situations
Probability cues	For example, soccer-team cue or state capital cue	For example, base rates of previous performance or average confidence in $N$ answers

*Note.* For illustration, questions of the Heidelberg-Bonn type are used. PMM = probabilistic mental model.

Table 1 summarizes the differences between PMMs that are implied by the two different tasks. Note that in our account, both confidences in a single event and judgments of frequency are explained by reference to experienced frequencies. However, these frequencies relate to different target variables and reference classes. We use this assumption to predict systematic differences between these kinds of judgments.

### *Adaptive PMMs and Representative Sampling*

A PMM is an inductive device that uses the “normal” life conditions in known environments as the basis for induction. How well does the structure of probability cues defined on  $R$  in a PMM represent the actual structure of probability cues in the environment? This question is also known as that of “proper cognitive adjustment” (Brunswik, 1964, p. 22). If the hierarchy of cues and their validities corresponds to that of the ecological validities, then the PMM is well adapted to a known environment. In Brunswik’s view, cue validities are learned by observing the frequencies of co-occurrences in an environment.

A large literature exists that suggests that (a) memory is often (but not always) excellent in storing frequency information from various environments and (b) the registering of event occurrences for frequency judgments is a fairly automatic cognitive process requiring very little attention or conscious effort (e.g., Gigerenzer, 1984; Hasher, Goldstein, & Toppino, 1977; Howell & Burnett, 1978; Zacks, Hasher, & Sanft, 1982). Hasher and Zacks (1979) concluded that frequency of occurrence, spatial location, time, and word meaning are among the few aspects of the environment that are encoded automatically and that encoding of frequency information is “automatic at least in part because of innate factors” (p. 360). In addition, Hintzman, Nozawa, and Irmscher (1982) proposed that frequencies are stored in memory in a nonnumerical analog mode.

Whatever the mechanism of frequency encoding, we use the following assumption for deriving our predictions: If subjects had repeated experience with a reference class, a target variable, and cues in their environment, we assume that cue validities correspond well to ecological validities. (This holds true for the average in a group of subjects, but individual idiosyncrasies in learning the frequency structure of the environment may occur.) This is a bold assumption made in ignorance of potential deviations between specific cue validities and ecological validities. If such deviations existed and were known, predictions by PMM theory could be improved. The assumption, however, derives support from both the literature on automatic frequency processing

and a large body of neo-Brunswikian research on the correspondence between ecological validities and cue utilization (the latter of which corresponds to our cue validities; e.g., Arkes & Hammond, 1986; K. Armelius, 1979; Brehmer & Joyce, 1988; MacGregor & Slovic, 1986).

Note that this adaptiveness assumption does not preclude that individuals (as well as the average subject) err. Errors can occur even if a PMM is highly adapted to a given environment. For instance, if an environment is changing or is changed in the laboratory by an experimenter, an otherwise well-adapted PMM may be suboptimal in a predictable way.

Brunswik's notion of "representative sampling" is important here. If a person experienced a representative sample of objects from a reference class, one can expect his or her PMM to be better adapted to an environment than if he or she happened to experience a skewed, unrepresentative sample.

Representative sampling is also important in understanding the relation between a PMM and the task. If a PMM is well adapted, but the set of objects used in the task (questions) is not representative of the reference class in the environment, performance in tasks will be systematically suboptimal.

To avoid confusion with terms such as *calibration*, we will use the term *adaptation* only when we are referring to the relation between a PMM and a corresponding environment—not, however, for the relation between a PMM and a task.

## Predictions

A concrete example can help motivate our first prediction. Two of our colleagues, K and O, are eminent wine tasters. K likes to make a gift of a bottle of wine from his cellar to Friend O, on the condition that O guesses what country or region the grapes were grown in. Because O knows the relevant cues, O can usually pick a region with some confidence. O also knows that K sometimes selects a quite untypical exemplar from his ample wine cellar to test Friend O's limits. Thus, for each individual wine, O can infer the probability that the grapes ripened in, say, Portugal as opposed to South Africa, with considerable confidence from his knowledge about cues. In the long run, however, O nevertheless expects the relative frequency of correct answers to be lower because K occasionally selects unusual items.

Consider tests of general knowledge, which share an important feature with the wine-tasting situation: Questions are selected to be somewhat difficult and sometimes misleading. This practice is common and quite reasonable for testing people's limits, as in the wine-tasting situation. Indeed, there is apparently not a single study on confidence in knowledge where a reference class has been defined and a representative (or random) sample of general-knowledge questions has been drawn from this population. For instance, consider the reference class "metropolis" and the geographical north-south location as the target variable. A question like "Which city is farther north? (a) New York or (b) Rome" is likely to appear in a general-knowledge test (almost everyone gets it wrong), whereas a comparison between Berlin and Rome is not.

The crucial point is that confidence and frequency judgments refer to different kinds of reference classes. A set of questions can be representative with respect to one reference class and, at the same time, selected with respect to the other class. Thus, a set of 50 general-knowledge questions of the city type may be representative for the reference class "sets of general-knowledge questions" but not for the reference class "cities in Germany" (because city pairs have been selected for being difficult or misleading). Asking for a confidence judgment summons up a PMM on the basis of the reference class "cities in Germany"; asking for a frequency judgment summons

up a PMM on the basis of the reference class “sets of general-knowledge questions.” The first prediction can now be stated.

1. *Typical general-knowledge tasks elicit both overconfidence and accurate frequency judgments.* By “typical” general-knowledge tasks we refer to a set of questions that is representative for the reference class “sets of general-knowledge questions.”

This prediction is derived in the following way: If (a) PMMs for confidence tasks are well adapted to an environment containing a reference class  $R$  (e.g., all cities in Germany) and (b) the actual set of questions is not representative for  $R$ , but selected for difficult pairs of cities, then confidence judgments exhibit overconfidence. Condition A is part of our theory (the simplifying assumption we just made), and Condition B is typical for the general-knowledge questions used in studies on confidence as well as in other testing situations.

If (a) PMMs for frequency-of-correct-answer tasks are well adapted with respect to an environment containing a reference class  $R'$  (e.g., the set of all general-knowledge tests experienced earlier), and (b) the actual set of questions is representative for  $R'$ , then frequency judgments are expected to be accurate. Again, Condition A is part of our theory, and Condition B will be realized in our experiments by using a typical set of general-knowledge questions.

Taken together, the prediction is that the same person will exhibit overconfidence when asked for the confidence that a particular answer is correct and accurate estimates when asked for a judgment of frequency of correct answers. This prediction is shown by the two points on the left side of Figure 4. This prediction cannot be derived from any of the previous accounts of overconfidence.

To introduce the second prediction, we return to the wine-tasting story. Assume that K changes his habit of selecting unusual wines from his wine cellar, and instead buys a representative sample of French red wines and lets O guess from what region they come. However, K does not tell O about the new sampling technique. O’s average confidence judgments will now be close to the proportion of correct answers. In the long run, O nevertheless expects the proportion of correct answers to be less, still assuming the familiar testing situation in which wines were selected, not randomly sampled. Thus, O’s frequency judgments will show underestimation.

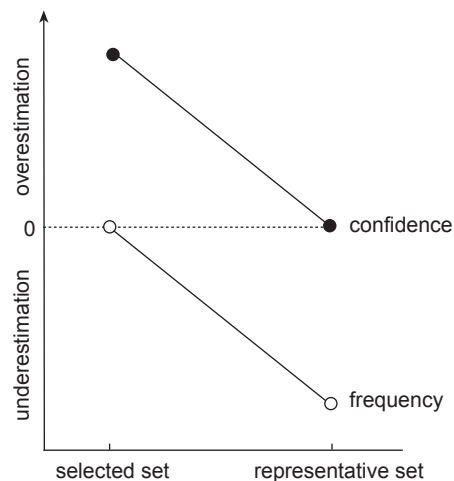


Figure 4. Predicted differences between confidence and frequency judgements (confidence-frequency effect).

Consider now a set of general-knowledge questions that is a random sample from a defined reference class in the subject's natural environment. We use the term *natural environment* to denote a knowledge domain familiar to the subjects participating in the study. This is a necessary (although not sufficient) condition to assume that PMMs are, on the average, well adapted. In the experiments reported herein, we used West German subjects and the reference class "all cities with more than 100,000 inhabitants in West Germany" (The study was conducted before the unification of Germany.) The second prediction is about this situation:

2. *If the set of general-knowledge tasks is randomly sampled from a natural environment, we expect overconfidence to be zero, but frequency judgments to exhibit underestimation.* Derivation is as before: If PMMs for confidence tasks are well adapted with respect to  $R$ , and the actual set of questions is a representative sample from  $R$ , then overconfidence is expected to disappear. If PMMs for frequency-of-correct-answers tasks are well adapted with respect to  $R'$ , and the actual set of questions is not representative for  $R'$  then frequency judgments are expected to be underestimations of true frequencies.

Again, this prediction cannot be derived from earlier accounts. Figure 4 shows Predictions 1 and 2. The predicted differences between confidence and frequency judgments is referred to as the *confidence-frequency effect*.

Testing these predictions also allows for testing the assumption of well-adapted PMMs for the confidence task. Assume that PMMs are not well adapted. Then a representative sample of city questions should not generate zero overconfidence but rather over- or underconfidence, depending on whether cue validities overestimate or underestimate ecological validities. Similarly, if PMMs for frequency judgments are not well adapted, frequency judgments should deviate from true frequencies in typical general-knowledge tasks. Independent of the degree of adaptation, however, the confidence-frequency effect should emerge, but the curves in Figure 4 would be transposed upward or downward.

We turn now to the standard way in which overconfidence has been demonstrated in previous research, comparing confidence levels with relative frequencies of correct answers at each confidence level. This standard comparison runs into a conceptual problem well-known in probability theory and statistics: A discrepancy between subjective probabilities in single events (i.e., the confidence that a particular answer is correct) and relative frequencies in the long run is not a bias in the sense of a violation of probability theory, as is clear from several points of view within probability theory. For instance, for a frequentist such as Richard von Mises (1928/1957), probability theory is about frequencies (in the long run), not about single events. According to this view, the common interpretation of overconfidence as a bias is based on comparing apples with oranges. What if that conceptual problem is avoided and, instead, the relative frequency of correct answers in each confidence category is compared with the estimated relative frequency in each confidence category? PMM theory makes an interesting prediction for this situation, following the same reasoning as for the frequency judgments in Predictions 1 and 2 (which were estimated frequency-of-correct answers in a series of  $N$  questions, whereas estimated relative frequencies in each confidence category are the concern here):

3. *Comparing estimated relative frequencies with true relative frequencies of correct answers makes overestimation disappear.* More precisely, if the set of general-knowledge questions is selected, over- or underestimation is expected to be zero; if the set is randomly sampled, underestimation is expected. Thus, PMM theory predicts that the distinction between confidence and relative frequency is psychologically real, in the sense that subjects do not believe that a confidence judgment of  $X\%$  implies a relative frequency of  $X\%$ , and vice versa. We know of no study on over-

confidence that has investigated this issue. Most have assumed instead that there is, psychologically, no difference.

Prediction 4 concerns the hard-easy effect, which says that overconfidence increases when questions get more difficult (e.g., Lichtenstein & Fischhoff, 1977). The effect refers to confidence judgments only, not to frequency judgments. On our account, the hard-easy effect is not simply a function of difficulty. Rather, it is a function of difficulty and a separate dimension, selected versus representative sampling. (Note that the terms *hard* and *easy* refer to the relative difficulty of two samples of items, whereas the terms *selected* and *representative* refer to the relation between one sample and a reference class in the person's environment.) PMM theory specifies conditions under which the hard-easy effect occurs, disappears, and is reversed. A reversed hard-easy effect means that overconfidence decreases when questions are more difficult.

In Figure 5, the line descending from H to E represents a hard-easy effect: Overconfidence in the hard set is larger than in the easy set. The important distinction (in addition to hard vs. easy) is whether a set was obtained by representative sampling or was selected. For instance, assume that PMMs are well adapted and that two sets of tasks differing in percentage correct (i.e., in difficulty) are both representative samples from their respective reference classes. In this case, one would expect all points to be on the horizontal zero-overconfidence line in Figure 5 and the hard-easy effect to be zero. More generally:

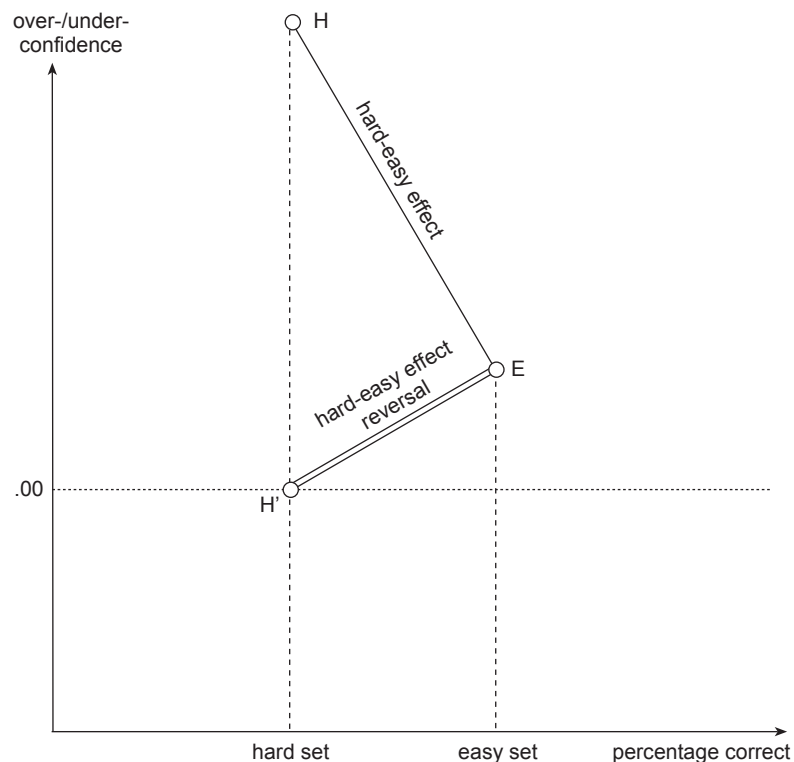


Figure 5. Predicted reversal of the hard-easy effect (H = hard; E = easy).

4. *If two sets, hard and easy, are generated by the same sampling process (representative sampling or same deviation from representative), the hard-easy effect is expected to be zero.* If sampling deviates in both the hard and the easy set equally from representative sampling, points will lie on a horizontal line parallel to the zero-overconfidence line.

Now consider the case that the easy set is selected from a corresponding reference class (e.g., general-knowledge questions), but the hard set is a representative sample from another reference class (denoted as  $H'$  in Figure 5). One then would predict a reversal of the hard-easy effect, as illustrated in Figure 5 by the double line from E to  $H'$ .

5. *If there are two sets, one is a representative sample from a reference class in a natural environment, the other is selected from another reference class for being difficult, but the representative set is harder than the selected set; then the hard-easy effect is reversed.* In the next section, Predictions 1, 2, and 3 are tested in two experiments; in the Explaining Anomalies in the Literature section, Predictions 4 and 5 are checked against results in the literature.

## Experiment 1

### *Method*

Two sets of questions were used, which we refer to as the *representative* and the *selected* set. The representative set was determined in the following way. We used as a reference class in a natural environment (an environment known to our subjects) the set of all cities in West Germany with more than 100,000 inhabitants. There were 65 cities (Statistisches Bundesamt, 1986). From this reference class, a random sample of 25 cities was drawn, and all pairs of cities in the random sample were used in a complete paired comparison to give 300 pairs. No selection occurred. The target variable was the number of inhabitants, and the 300 questions were of the following kind: "Which city has more inhabitants? (a) Solingen or (b) Heidelberg." We chose city questions for two reasons. First, and most important, this content domain allowed for a precise definition of a reference class in a natural environment and for random sampling from this reference class. The second reason was for comparability. City questions have been used in earlier studies on overconfidence (e.g., Keren, 1988; May, 1987).

In addition to the representative set, a typical set of general-knowledge questions, as in previous studies, was used. This selected set of 50 general-knowledge questions was taken from an earlier study (Angele et al., 1982). Two examples are "Who was born first? (a) Buddha or (b) Aristotle" and "When was the zipper invented? (a) before 1920 or (b) after 1920."

After each answer, the subject gave a confidence judgment (that this particular answer was correct). Two kinds of frequency judgments were used. First, after each block of 50 questions, the subject estimated the number of correct answers among the 50 answers given. Because there were 350 questions, every subject gave seven estimates of the number of correct answers. Second, after the subjects answered all questions, they were given an enlarged copy of the confidence scale used throughout the experiment and were asked for the following frequency judgment: "How many of the answers that you classified into a certain confidence category are correct? Please indicate for every category your estimated relative frequency of correct answers."

In Experiment 1, we also introduced two of the standard manipulations in the literature. The first was to inform and warn half of our subjects of the overconfidence effect, and the second was to offer half of each group a monetary incentive for good performance. Both are among a list of "debiasing" methods known as being relatively ineffective (Fischhoff, 1982), and both contrib-

uted to the view that overconfidence is a robust phenomenon. If PMM theory is correct, the magnitude of effects resulting from the two manipulations—confidence versus frequency judgment and selected versus representative sampling—should be much larger than those resulting from the “debiasing” manipulations.

*Subjects.* Subjects were 80 students (43 men and 37 women) at the University of Konstanz who were paid for participation. Eighty-five percent of them grew up in the state of Baden-Württemberg, so the group was fairly homogeneous (knowledge about city populations often depends on the rater’s geographical location). Subjects were tested in small groups of a maximum of 12 persons.

*Design and procedure.* This was a  $2 \times 2 \times 2$  design with representative-selected set varied within subjects, and information-no information about overconfidence and monetary incentive-no incentive as independent variables varied between subjects. Half of the subjects answered the representative set first; the other half, the selected set. Order of questions was determined randomly in both sets.

The confidence scale consisted of seven categories, 50%, 51%–60%, 61%–70%, 71%–80%, 81%–90%, 91%–99%, and 100% confident. The 50%- and 100%-confidence values were introduced as separate categories because previous research showed that subjects often tend to use these particular values. Subjects were told first to mark the alternative that seemed to be the correct one, and then to indicate with a second cross their confidence that the answer was correct. If they only guessed, they should cross the 50% category; if they were absolutely certain, they should cross the 100% category. We explained that one of the alternatives was always correct. In the information condition, subjects received the following information: “Most earlier studies found a systematic tendency to overestimate one’s knowledge; that is, there were many fewer answers correct than one would expect from the confidence ratings given. Please keep this warning in mind.” In the incentive condition, subjects were promised 20 German marks (or a bottle of French champagne), in addition to the payment that everyone received (7.50 marks), for the best performance in the group.

To summarize, 350 questions were presented, with a confidence judgment after each question, a frequency judgment after each 50 questions, and a judgment of relative frequencies of correct answers in each confidence category at the end.

For comparison with the literature on calibration, we used the following measure:

$$\text{over- or underconfidence} = \frac{1}{n} \sum_{i=1}^I n_i (p_i - f_i) = \bar{p} - \bar{f}$$

where  $n$  is the total number of answers,  $n_i$  is the number of times the confidence judgment  $p_i$  was used, and  $f_i$  is the relative frequency of correct answers for all answers assigned confidence  $p_i$ .  $I$  is the number of different confidence categories used ( $I = 7$ ), and  $\bar{p}$  and  $\bar{f}$  are the overall mean confidence judgment and percentage correct, respectively. A positive difference is called overconfidence. For convenience, we report over- and underconfidence in percentages ( $\times 100$ ).

## Results

*Prediction 1.* PMM theory predicts that in the selected set (general-knowledge questions), people show overestimation in confidence judgments (overconfidence) and, simultaneously, accurate frequency judgments.

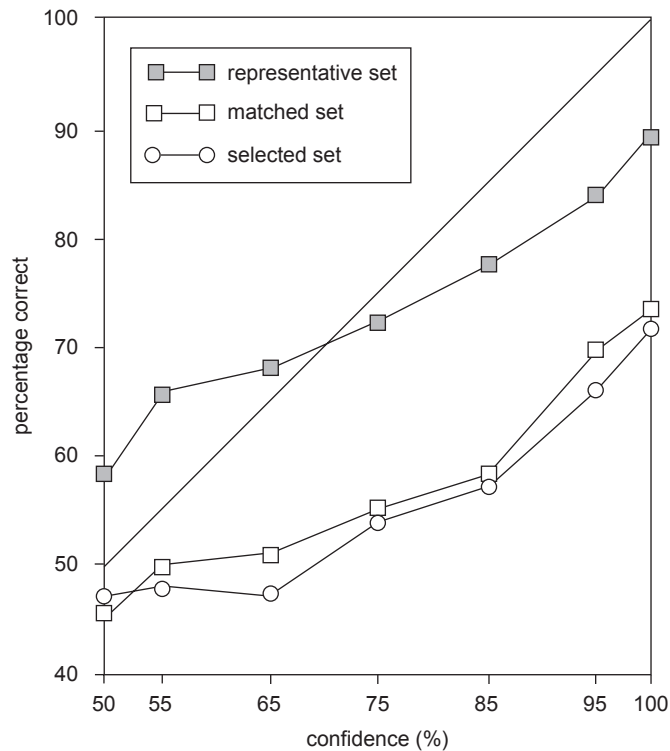


Figure 6. Calibration curves for the selected set (open circles), representative set (black squares), and matched set (open squares).

The open-circle curve in Figure 6 shows the relation between judgments of confidence and true relative frequency of correct answers in the selected set—that is, the set of mixed general-knowledge questions.<sup>4</sup> The relative frequency of correct answers (averaged over all subjects) was 72.4% in the 100%-confidence category, 66.3% in the 95% category, 58.0% in the 85% category, and so on. The curve is far below the diagonal (calibration curve) and similar to the curves reported by Lichtenstein, Fischhoff, and Phillips (1982, Figure 2). It replicates and demonstrates the well-known overconfidence effect. Percentage correct was 52.9, mean confidence was 66.7, and overconfidence was 13.8.

Subjects' frequency judgments, however, are fairly accurate, as Table 2 (last row) shows. Each entry is averaged over the 20 subjects in each condition. For instance, the figure  $-1.75$  means that, on average, subjects in this condition underestimated the true number of correct answers by 1.75. Averaged across the four conditions, we get  $-1.2$ , which means that subjects missed the true frequency by an average of only about 1 correct answer in the set of 50 questions. Quite accurate frequency judgments coexist with overconfidence. The magnitudes of this confidence–frequency effect found is shown in Figure 7 (left side). PMM theory predicts this systematic difference be-

<sup>4</sup> In Figure 6, we have represented the confidence category (91%–99%) by 95%, and similarly with the other categories. This choice can be criticized because numerical judgments of confidence often cluster around specific values in an interval. (If there is a difference, however, we may expect that it affects the three curves in a similar way, without altering the differences between curves.) In Experiment 2, we used precise values instead of these intervals.

Table 2  
*Mean Differences Between Estimated and True Frequencies of Correct Answers*

Set	No information– no incentive	Incentive only	Information only	Information and incentive
Representative				
1–50	–9.94	–9.42	–8.80	–8.74
51–100	–9.50	–10.37	–11.95	–11.25
101–150	–9.88	–10.89	–10.85	–9.90
151–200	–6.67	–6.70	–9.35	–5.90
201–250	–9.79	–9.84	–7.95	–5.25
251–300	–9.47	–10.84	–9.40	–9.05
Average	–9.21	–9.68	–9.72	–8.35
Selected	–1.75	–0.60	–2.65	0.30

*Note.* Negative signs denote underestimation of true number of correct answers.

tween confidence and frequency judgments, within the same person and the same general-knowledge questions.

*Prediction 2.* PMM theory predicts that in the representative set (city questions) people show zero overconfidence and, at the same time, underestimation in frequency judgments.

The solid-square curve in Figure 6 shows the relation between confidence and percentage correct in the representative set—that is, the city questions. For instance, percentage correct in the 100%–confidence category was 90.8%, instead of 72.4%. Overconfidence disappeared (–0.9%). Percentage correct and mean confidence were 71.7 and 70.8, respectively.

The confidence curve for the representative set is similar to a regression curve for the estimation of relative frequencies by confidence, resulting in underconfidence in the left part of the confidence scale, overconfidence in the right, and zero overconfidence on the average.

Table 2 shows the differences between estimated and true frequencies for each block of 50 items and each of the conditions, respectively. Again, each entry is averaged over the 20 subjects in each condition. For instance, subjects who were given neither information nor incentive underestimated their true number of correct answers by 9.94 (on the average) in the first 50 items of the representative set. Table 2 shows that the values of the mean differences were fairly stable over the six subsets, and, most important, they are, without exception, negative (i.e., underestimation). For each of the 24 cells (representative set), the number of subjects with negative differences (underestimation) was compared with the number of positive differences (overestimation) by sign tests, and all 24 *p* values were smaller than .01.

The following is an illustration at the individual level: Subject 1 estimated 28, 30, 23, 25, 23, and 23, respectively, for the six subsets, compared with 40, 38, 40, 36, 35, and 32 correct solutions, respectively. An analysis of individual judgments confirmed average results. Among the 80 subjects, 71 underestimated the number of correct answers, whereas only 8 subjects overestimated it (frequency judgments were missing for 1 subject). Incidentally, 7 of these 8 subjects were male. In the selected set, for comparison, 44 subjects underestimated and 35 subjects overestimated the number of correct answers, and 1 subject got it exactly right.

We have attributed the emergence and disappearance of overconfidence to selection versus use of a representative set. One objection to this analysis is that the difference between the open-circle and the solid-square curve in Figure 6 is confounded with a difference in the content of both sets. The selected set includes a broad range of general-knowledge questions, whereas the domain of the representative set (cities) is necessarily more restricted. To check for this possible

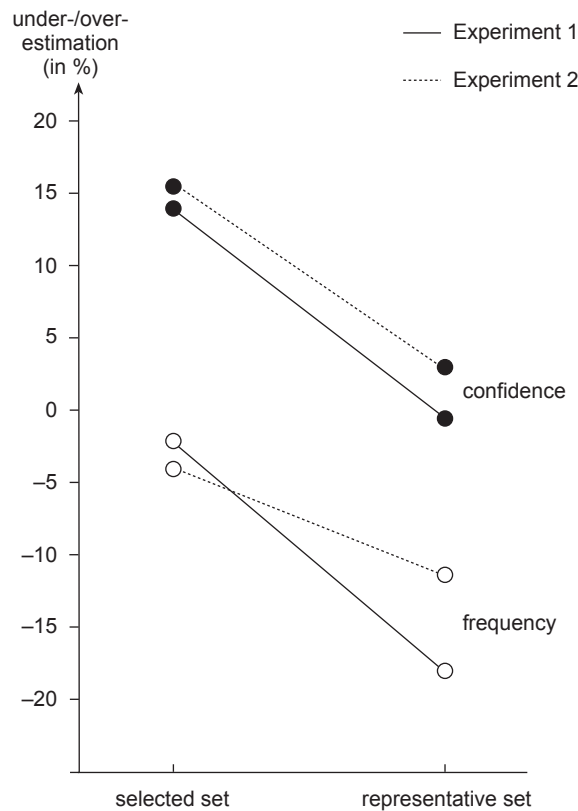
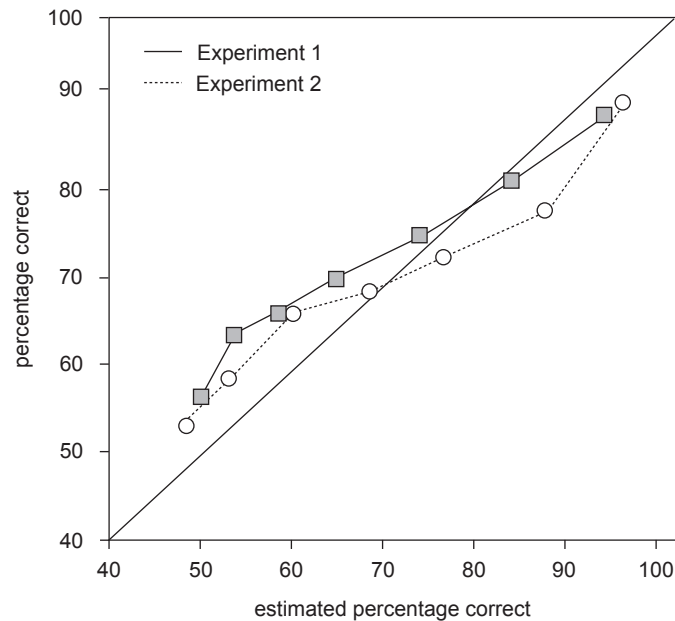


Figure 7. Confidence-frequency effect in the representative and selected set. (Solid lines show results of Experiment 1, dotted lines show those of Experiment 2. Frequency judgments are long-run frequencies,  $N = 50$ .)

confound, we determined the item difficulties for each of the 50 general-knowledge questions and selected a subset of 50 city questions that had the same item difficulties. If the difference in Figure 6 is independent of content, but results from the selection process, this “matched” subset of city questions should generate the same calibration curves showing overconfidence as the selected set of general-knowledge questions did. Figure 6 shows that this is the case (open-square curve). Both content domains produce the same results if questions are selected.

To summarize, in the representative set, overestimation disappears in confidence judgments, and zero-overconfidence coexists with frequency judgments that show large underestimation. Results confirm Prediction 2. Figure 7 (right side) shows the magnitude of the confidence-frequency effect found. No previous theory of confidence can predict the results depicted in Figure 7.

*Prediction 3.* PMM theory predicts that overestimation will disappear if the relative frequencies of correct answers (percentage correct) in each confidence category is compared with the estimated relative frequencies. Because subjects estimated percentage correct for all confidence judgments—that is, including both the selected and the representative set—we expect not only that overestimation will disappear (the prediction from the selected set) but also that it will turn into underestimation (the prediction from the representative set).



*Figure 8.* Calibration curves for judgments of percentage correct in confidence categories. (Solid lines show results of Experiment 1, dotted lines show those of Experiment 2. Values are averaged across both sets of questions.)

The solid line in Figure 8 shows the results for Experiment 1: Estimated relative frequencies are well calibrated and show underestimation in five out of seven confidence categories. Overestimation of one's knowledge disappears. The only exception is the 100%-confidence category. The latter is the confidence category that contains all solutions by local MMs, and errors in memory or elementary logical operations may account for the difference. Figure 8 is a "frequentist" variant of the calibration curve of Figure 6. Here, true percentage correct is compared with estimated percentage correct, rather than with confidence. For instance, in the 100%-confidence category, true and estimated percentage correct were 88.8% and 93.0%, respectively.

Averaged across experimental conditions, the ratio between estimated frequency in the long run and confidence value is fairly constant, around .87, for confidence ratings between 65% and 95%. It is highest in the extreme categories (see Table 3).

To summarize, subjects explicitly distinguished between confidence in single answers and the relative frequency of correct answers associated with a confidence judgment. This result is implied by PMM theory, according to which different reference classes are cued by confidence and frequency tasks. As stated in Prediction 3, overestimation disappeared. However, the magnitude of underestimation was not, as might be expected, as pronounced as in the frequency judgments dealt with in Predictions 1 and 2. Except for this finding, results conformed well to Prediction 3. Note that no previous theory of confidence in knowledge we are aware of makes this conceptual distinction and that prediction. Our results contradict much of what has been assumed about how the untutored mind understands the relation between confidence and relative frequency of correct answers.

Table 3  
*Estimated and True Percentage Correct in Each Confidence Category*  
*(Summarized Over the Representative and the Selected Sets)*

Confidence category	No. of confidence judgments	% correct		Over-/under estimation
		Estimated	True	
100	5,166	93.0	88.8	4.2
91–99	1,629	82.7	81.6	1.1
81–90	2,534	73.1	74.6	–1.5
71–80	2,950	64.3	70.1	–5.8
61–70	3,506	57.3	65.6	–8.3
51–60	4,036	53.7	63.3	–9.6
50	8,178	49.8	56.3	–6.5
$\Sigma$ or <i>M</i>	27,999	64.8	69.1	–4.2

*Information about overconfidence and monetary incentive.* Mean confidence judgments were indistinguishable between subjects informed about overconfidence and those uninformed. None of seven *t* tests, one for each confidence category, resulted in *p* values smaller than .05. If a monetary incentive was announced, overconfidence was more pronounced with incentive than without incentive in five categories (65%–100%) and less in the 50% category (all *ps* < .05), with an average increase of 3.6%.

The monetary incentive effect resulted from the incentive/no-information group, in which confidence judgments were higher than in all three other groups (but we found the same percentage correct in all groups). One reason for this interaction could be that we did not specify in the instructions a criterion for best performance. If warned of overconfidence, subjects could easily infer that the incentive was for minimizing overconfidence. If not warned, at least some subjects could also have attempted to maximize percentage correct. None of these attempts, however, was successful, consistent with PMM theory and earlier studies (e.g., Fischhoff, Slovic, & Lichtenstein, 1977). The effort to raise the percentage correct seems to have raised confidence instead, an outcome that cannot be accounted for by PMM theory. The size of this effect, however, was small compared with both the confidence-frequency effect and that of selected versus representative sampling.

To summarize, neither warning of overconfidence nor associated monetary incentive decreased overconfidence or increased percentage correct, replicating earlier findings that knowledge about overconfidence is not sufficient to change confidence. An incentive that subjects seem to have interpreted as rewarding those who maximize the percentage correct, however, increased confidence.

*Order of presentation and sex.* Which set (representative vs. selected) was given first had no effect on confidences, neither in Experiment 1 nor in Experiment 2. Arkes, Christensen, Lai, and Blumer (1987) found an effect of the difficulty of one group of items on the confidence judgments for a second when subjects received feedback for their performance in the first set. In our experiment, however, no feedback was given. Thus, subjects had no reason to correct their confidence judgments, such as by subtracting a constant value. Sex differences in degree of overconfidence in knowledge have been claimed by both philosophy and folklore. Our study, however, showed no significant differences between the sexes in either overconfidence or calibration, in either Experiment 1 or in Experiment 2. (The men's confidence judgments were on the average 5% higher than women's, but so was their percentage correct. This replicates Lichtenstein and Fischhoff's, 1981, findings about students at the University of Oregon.)

To summarize, as predicted by PMM theory, we can experimentally make overconfidence (overestimation) appear, disappear, and invert. Experiment 1 made our subjects consistently switch back and forth among these responses. The key to this finding is a pair of concepts that have been neglected by the main previous explanations of confidence in one's knowledge—confidence versus frequency judgment and representative versus selected sampling.

## Experiment 2

We tried to replicate the facts and test several objections. First, to strengthen the case against this theory, we instructed the subjects both verbally and in written form that confidence is subjective probability, and that among all cases where a subjective probability of X% was chosen, X% of the answers should be correct. Several authors have argued that such a frequentist instruction could enhance external calibration or internal consistency (e.g., Kahneman & Tversky, 1982; May, 1987). According to PMM theory, however, confidence is already inferred from frequency (with or without this instruction)—but from frequencies of co-occurrences between, say, number of inhabitants and several cues, and not from base rates of correct answers in similar testing situations (see Table 1). Thus, in our view, the preceding caution will be ineffective because the base rate of correct answers is not a probability cue that is defined on a reference class such as cities in Germany.

Second, consider the confidence-frequency effect. We have shown that this new effect is implied by PMM theory. One objection might be that the difference between confidence and frequency judgments is an artifact of the response function, just as overconfidence has sometimes been thought to be. Consider the following interpretation of overconfidence. If (a) confidence is well calibrated but (b) the response function that transforms confidence into a confidence judgment differs from an identity function, then (c) overconfidence or underconfidence “occurs” on the response scale. Because an identity function has not been proven, Anderson (1986), for instance, denoted the overconfidence effect and the hard-easy effect as “largely meaningless” (p. 91): They might just as well be response function artifacts.

A similar objection could be made against the interpretation of the confidence-frequency effect within PMM theory. Despite the effect's stability across selected and representative sets, it may just reflect a systematic difference between response functions for confidence and frequency judgments. This conjecture can be rephrased as follows: If (a) the difference between “internal” confidence and frequency impression is zero, but (b) the response functions that transform both into judgments differ systematically, then (c) a confidence-frequency effect occurs on the response scales. We call this the *response-function conjecture*.

How can this conjecture be tested? According to PMM theory, the essential basis on which both confidence and frequency judgments are formed is the probability cues, not response functions. We assumed earlier that frequency judgments are based mainly on base rates of correct answers in a reference class of similar general-knowledge test situations. If we make another cue available, then frequency judgments should change. In particular, if we make the confidence judgments more easily retrievable from memory, these can be used as additional probability cues, and the confidence-frequency effect should decrease. This was done in Experiment 2 by introducing frequency judgments in the short run, that is, frequency judgments for a very small number of questions. Here, confidence judgments can be more easily retrieved from memory than they could in the long run. Thus, if PMM theory is correct, the confidence-frequency effect should decrease in the short run. If the issue were, however, different response functions, then the availability of confidence judgments should not matter because confidence and frequency im-

pression are assumed to be identical in the first place. Thus, if the conjecture is correct, the confidence-frequency effect should be stable.

In Experiment 2, we varied the length  $N$  of a series of questions from the long run condition  $N = 50$  in Experiment 1 to the smallest possible short run of  $N = 2$ .

Third, in Experiment 1 we used a response scale ranging from 50% to 100% for confidence judgments but a full-range response scale for frequency judgments ranging from 0 to 50 correct answers (which corresponds to 0% to 100%). Therefore one could argue that the confidence-frequency effect is an artifact of the different ranges of the two response scales. Assume that (a) there is no difference between internal confidence and frequency, but (b) because confidence judgments are limited to the upper half of the response scale, whereas frequency judgments are not, (c) the confidence-frequency effect results as an artifact of the half-range response scale in confidence judgments. We refer to this as the *response-range conjecture*. It can be backed up by at least two hypotheses.

1. Assume that PMM theory is wrong and subjects indeed use base rates of correct answers as a probability cue for confidence in single answers. Then confidence should be considerably lower. If subjects anticipate misleading questions, even confidences lower than 50% are reasonable to expect on this conjecture. Confidences below 50%, however, cannot be expressed on a scale with a lower boundary at 50%, whereas they can at the frequency scale. Effects of response range such as those postulated in range-frequency theory (Parducci, 1965) or by Schönemann (1983) may enforce the distorting effect of the half-range format. In this account, both the overconfidence effect and the confidence-frequency effect are generated by a response-scale effect. With respect to overconfidence, this conjecture has been made and has claimed some support (e.g., May, 1986, 1987; Ronis & Yates, 1987). We call this the *base rate hypothesis*.

2. Assume that PMM theory is wrong in postulating that choice and confidence are essentially one process and that the true process is a temporal sequence: choice, followed by search for evidence, followed by confidence judgment. Koriat et al. (1980), for instance, proposed this sequence. Assume further, contrary to Koriat, that the mind is “Popperian,” searching for disconfirming rather than for confirming evidence to determine the degree of “corroboration” of an answer. If the subject is successful in retrieving disconfirming evidence from memory, but is not allowed to change the original answer, confidence judgments less than 50% will result. Such disconfirmation strategies, however, can hardly be detected using a 50%–100% format, whereas they could in a full-scale format. We call this the *disconfirmation strategy hypothesis*.

To test the response-range conjecture, half of the subjects in Experiment 2 were given full-range response scales, whereas the other half received the response scales used in Experiment 1.

## Method

*Subjects.* Ninety-seven new subjects at the University of Konstanz (not enrolled in psychology) were paid for participation. There were 59 male and 38 female subjects. As in Experiment 1, subjects were tested in small groups of no more than 7 subjects.

*Design and procedure.* This was a  $4 \times 2 \times 2$  design, with length of series (50, 10, 5, and 2) and response scale (half range vs. full range) varied between subjects and type of knowledge questions (selected vs. representative set) varied within subjects.

The procedure and the materials were like that in Experiment 1, except for the following. We used a new random sample of 21 (instead of 25) cities. This change decreased the number of questions in the representative set from 300 to 210. As mentioned earlier, we explicitly instructed

the subjects to interpret confidences as frequencies of correct answers: “We are interested in how well you can estimate subjective probabilities. This means, among all the answers where you give a subjective probability of X%, there should be X% of the answers correct.” This calibration instruction was orally repeated and emphasized to the subjects.

The response scale contained the means (50%, 55%, 65%, ..., 95%, 100%) of the intervals used in Experiment 1 rather than the intervals themselves to avoid the problematic assumption that means would represent intervals. Endpoints were marked *absolutely certain that the alternative chosen is correct* (100%), *both alternatives equally probable* (50%), and, for the full-range scale, *absolutely certain that the alternative chosen is incorrect* (0%). In the full-range scale, one reason for using confidences between 0% and 45% was explained in the following illustration: “If you think after you have made your choice that you would have better chosen the other alternative, do not change your choice, but answer with a probability smaller than 50%.”

After each set of  $N = 50$  (10, 5, or 2) answers, subjects gave a judgment of the number of correct answers. After having completed  $50 + 210 = 260$  confidence judgments and 5, 26, 52, or 130 frequency judgments (depending on the subject’s group), subjects in both response-scale conditions were presented the same enlarged copy of the 50%–100% response scale and asked to estimate the relative frequency of correct answers in each confidence category.

## *Results*

*Response-range conjecture.* We tested the conjecture that the systematic difference in confidence and frequency judgments stated in Predictions 1 and 2 (confidence-frequency effect) and shown in Experiment 1 resulted from the availability of only a limited response scale for confidence judgments (50% to 100%).

Forty-seven subjects were given the full-range response scale for confidence judgments. Twenty-two of these never chose confidences below 50%; the others did. The number of confidence judgments below 50% was small. Eleven subjects used them only once (in altogether 260 judgments), 5 did twice, and the others 3 to 7 times. There was one outlier, a subject who used them 67 times. In total, subjects gave a confidence judgment smaller than 50% for only 1.1% of their answers (excluding the outlier: 0.6%). If the response-range conjecture had been correct, subjects would have used confidence judgments below 50% much more frequently.

In the representative set, overconfidence was 3.7% ( $SE_M = 1.23$ ) in the full-range scale condition and 1.8% ( $SE_M = 1.15$ ) in the half-range condition. In the selected set, the corresponding values were 14.4 ( $SE_M = 1.54$ ) and 16.4 ( $SE_M = 1.43$ ). Averaging all questions, we got slightly larger overconfidence in the full-range condition (mean difference = 1.2). The response-range conjecture, however, predicted a strong effect in the opposite direction. Frequency judgments were essentially the same in both conditions. Hence, the confidence-frequency effect can also be demonstrated when both confidence and frequency judgments are made on a full-range response scale.

To summarize, there was (a) little use of confidences below 50% and (b) no decrease of overconfidence in the full-range condition. These results contradict the response-range conjecture.

A study by Ronis and Yates (1987) seems to be the only other study that has compared the full-range and the half-range format in two-alternative choice tasks, but it did not deal with frequency judgments. These authors also reported that only about half their subjects used confidence judgments below 50%, although they did so more frequently than our subjects. Ronis and Yates concluded that confidences below 50% had only a negligible effect on overconfidence and

calibration (pp. 209–211). Thus, results in both studies are consistent. The main difference is that Ronis and Yates seem to consider only “failure to follow the instructions” and “misusing the probability scale” (p. 207) as possible explanations for confidence judgments below 50%. In contrast, we argue that there are indeed plausible cognitive mechanisms—the base rate and disconfirmation strategy hypotheses—that imply these kind of judgments, although they would contradict PMM theory.

Both Experiment 2 and the Ronis and Yates (1987) study do not rule out, however, a more fundamental conjecture that is difficult to test. This argument is that internal confidence (not frequency) takes a verbal rather than a numerical form and that it is distorted on any numerical probability rating scale, not just on a 50%–100% response scale. Zimmer (1983, 1986) argued that verbal expressions of uncertainty (such as “highly improbably” and “very likely”) are more realistic, more precise, and less prone to overconfidence and other so-called judgmental biases than are numerical judgments of probability. Zimmer’s fuzzy-set modeling of verbal expressions, like models of probabilistic reasoning that dispense with the Kolmogoroff axioms (e.g., Cohen, 1989; Kyburg, 1983; Shafer, 1978), remains a largely unexplored source of alternative accounts of confidence.

For the remaining analysis, we do not distinguish between the full-range and the half-range response format. For combining the data, we recoded answers like “alternative *a*, 40% confident” as “alternative *b*, 60% confident,” following Ronis and Yates (1987).

*Predictions 1 and 2: Confidence-frequency effect.* The question is whether the confidence-frequency effect can be replicated under the explicit instruction that subjective probabilities should be calibrated to frequencies of correct answers in the long run. Calibration curves in Experiment 2 were similar to those in Figure 6 and are not shown here for this reason. Figure 7 shows that the confidence-frequency effect replicates. In the selected set, mean confidence was 71.6%, and percentage correct was 56.2. Mean estimated number of correct answers (transformed into percentages) in the series of  $N = 50$  was 52.0%. As stated in Prediction 1, overconfidence in single answers coexists with fairly accurate frequency judgments, which once again show slight underestimation.

In the representative set, mean confidence was 78.1% and percentage correct was 75.3%.<sup>5</sup> Mean estimated number of correct answers per 50 answers was 63.5%. As forecasted in Prediction 2, overconfidence largely disappeared (2.8%), and frequency judgments showed underestimation (–11.8%).

An individual analysis produced similar results. The confidence-frequency effect (average confidence higher than average frequency judgment) held for 82 (83) subjects in the selected (representative) set (out of 97). Answering the selected set, 92 respondents showed overconfidence, and 5 showed underconfidence. In the representative set, however, 60 exhibited overconfidence and 37, underconfidence.

*Prediction 3: Estimated percentage correct in confidence categories.* After the subjects answered the 260 general-knowledge questions, they were asked what percentage they thought they had correct in each confidence category. As shown by the dashed line in Figure 8, results replicated well. Average estimated percentage correct differed again from confidence and was close to the actual percentage correct.

<sup>5</sup> Confidence and percentage correct are averaged across all four conditions (series length) because these do not differ systematically among conditions. For comparison, the corresponding values for confidence and percentage correct in the  $N = 50$  condition are 71.0 and 56.8 in the selected set and 79.2 and 74.5 in the representative set.

Despite the instruction not to do so, our subjects still distinguished between a specific confidence value and the corresponding percentage of correct responses. Therefore *confidence* and *hypothesized percentage correct* should not be used as synonyms (e.g., Dawes, 1980, pp. 331–345). As suggested by this experiment, an instruction alone cannot override the cognitive processes at work.

In the 100%-confidence category, for instance, 67 subjects gave estimates below 100%. In a postexperimental interview, we pointed out to them that these judgments imply that they assumed they had not followed the calibration instruction. Most subjects explained that in each single case, they were in fact 100% confident. But they also knew that, in the long run, some answers would nonetheless be wrong, and they did not know which ones. Thus, they did not know which of the 100% answers they should correct. When asked how they made the confidence judgments, most subjects answered by giving examples of probability cues, such as “I know that this city is located in the Ruhrgebiet, and most cities there are rather large.” Interviews provided evidence for several probability cues, but no evidence that base rate expectations, as reported in frequency judgments, were also used in confidence judgments.

*Response-function conjecture: Frequency judgments in the short and long runs.* We tested the conjecture that the confidence-frequency effect stated in Predictions 1 and 2 and shown in Experiment 1 might be due to different response functions for confidence and frequency judgments, rather than to different cognitive processes as postulated by PMM theory. If the conjecture were true, the availability of confidence judgments in the short run should not change the confidence-frequency effect (see the previous discussion).

Contrary to the response-function conjecture, the length of series showed a significant effect on the judgments of frequency of correct answers in each series ( $p = .025$ ) as well as on the difference between judged and true frequency ( $p = .012$ ). Figure 9 shows the extent of the disappearance of the confidence-frequency effect in the short run. The curve shows that the effect decreased from  $N = 50$  to  $N = 2$ , averaged across both sets of items. The decrease was around 12%, an amount similar in the selected set (from 18.9% to 6.9%) and in the representative set (from 15.7% to 3.3%).

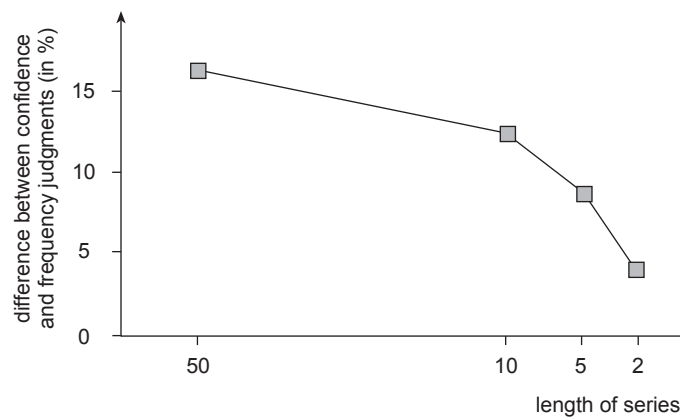


Figure 9. Decrease of the confidence-frequency effect in short runs ( $N = 50, 10, 5,$  and  $2$ ). (Values are differences between mean confidence and estimated percentage correct in a series of length  $N$ . Values are averaged across all questions.)

As would be expected from both the response-function conjecture and PMM theory, an analysis of variance over all 260 questions showed no significant effect of length of series (short vs. long runs) on either confidence judgment ( $p = .39$ ) or number of correct answers ( $p = .40$ ). (Similar results were obtained when the selected and representative sets were tested separately.)

The breakdown of the confidence-frequency effect in the short run is inconsistent with the objection that the effect can be reduced to a systematic difference in response functions. This result is, however, consistent with the notion that the shorter the run, the more easily are confidence judgments available from memory, and, thus, the more they can be used as probability cues for the true number of correct answers.

### *Discussion*

Our starting point was the overconfidence effect, reported in the literature as a fairly stable cognitive illusion in evaluating one's general knowledge and attributed to general principles of memory search, such as confirmation bias (Koriat et al., 1980), to general motivational tendencies such as fear of invalidity (Maysseles & Kruglanski, 1987), to insensitivity to task difficulty (see von Winterfeldt & Edwards, 1986, p. 128), and to wishful thinking and other "deficits" in cognition, motivation, and personality. Our view, in contrast, proposes that one evaluates one's knowledge by probabilistic mental models. In our account, the main deficit of most cognitive and motivational explanations is that they neglect the structure of the task and its relation to the structure of a corresponding environment known to the subjects. If people want to search for confirming evidence or to believe that their answers are more correct than they are because of some need, wish, or fear, then overestimation of accuracy should express itself independently of whether they judge single answers or frequencies, a selected or representative sample of questions, and hard or easy questions.

Our experiments also do not support the explanation of overconfidence and the hard-easy effect by assuming that subjects are insensitive to task difficulty: In frequency tasks we have shown that subjects' judgments of their percentage correct in the long run are in fact close to actual percentage correct, although confidences are not. Overconfidence does not imply that subjects are not aware of task difficulty. At least two more studies have shown that estimated percentage correct can correspond closely to true percentage correct in general-knowledge tasks. Allwood and Montgomery (1987) asked their subjects to estimate how difficult each of 80 questions was for their peers and found that difficulty ratings ( $M = 57\%$ ) were more realistic (percentage correct = 61%) than confidence judgments ( $M = 74\%$ ). May (1987) asked her subjects to estimate their percentage of correct answers after they completed an experiment with two-alternative questions. She found that judgments of percentage correct accorded better with the true percentage correct than did confidences.

On our account, overconfidence results from one of two causes, or both: (a) A PMM for a task is not properly adapted to a corresponding environment (e.g., cue validities do not correspond to ecological validities), or (b) the set of objects used is not a representative sample from the corresponding reference class in the environment but is selected for difficulty. If a is the true cause, using a representative sample from a known environment should not eliminate overconfidence. If b is true, it should. In both experiments, overconfidence in knowledge about city populations was eliminated, as implied by b. Thus, experimental results are consistent with both PMM theory and the assumption that individual PMMs are on the average well adapted to the

city environment we used.<sup>6</sup> Overconfidence resulted in a set of questions that was selected for difficulty. Underconfidence, conversely, would result from questions selected to be easy.

The foregoing comments do not mean that overestimation of knowledge is just an artifact of selected questions. If it were, then judgments of frequency of correct answers should show a similar degree of overestimation. What we have called the confidence-frequency effect shows that this is not the case.

Several authors have proposed that judgments in the frequency mode are more accurate, realistic, or internally consistent than probabilities for single events (e.g., Teigen, 1974, p. 62; Tversky & Kahneman, 1983). Our account is different. PMM theory states conditions under which mean judgments of confidence are systematically larger than judgments of relative frequency. PMM theory does not, however, imply that frequency judgments are generally better calibrated. On the contrary, frequency judgments may be miscalibrated for the same reasons as confidence judgments. The set of tasks may not be representative for the reference class from which the inferences are made.

The experimental control of overestimation—how to make overestimation appear, disappear, and invert—gives support to PMM theory. These predictions, however, do not exhaust the inferences that can be derived from PMM theory.

### *Explaining Anomalies in the Literature*

In this section, we explain a series of apparently inconsistent findings and integrate these into PMM theory.

*Ronis and Yates (1987).* We have mentioned that the Ronis and Yates' (1987) study is the only other study that tested a full-range response scale for two-alternative tasks. The second purpose of that study was to compare confidence judgments in situations where the subject knows that the answers are known to the experimenter (general-knowledge questions) with outcomes of upcoming basketball games, where answers are not yet known. In all three (response-scale) groups, percentage correct was larger for general-knowledge questions than for basketball predictions. Given this result, what would current theories predict about overconfidence? The insensitivity hypothesis proposes that people are largely insensitive to percentage correct (see von Winterfeldt & Edwards, 1986, p. 128). This implies that overconfidence will be larger in the more difficult (hard) set: the hard-easy effect. (The confirmation bias and motivational explanations are largely mute on the difficulty issue.) PMM theory, in contrast, predicts that overconfidence will be larger in the easier set (hard-easy effect reversal, see Prediction 5) because general-knowledge questions (the easy set) were selected and basketball predictions were not; only with clairvoyance could one select these predictions for percentage correct.

In fact, Ronis and Yates (1987) reported an apparent anomaly: three hard-easy effect reversals. In all groups, overconfidence was larger for the easy general-knowledge questions than for the hard basketball predictions (Figure 10). Ronis and Yates seem not to have found an explanation for these reversals of the hard-easy effect.

*Koriat et al. (1980).* Experiment 2 of Koriat et al.'s (1980) study provided a direct test of the confirmation bias explanation of overconfidence. The explanation is this: (a) Subjects first choose

---

<sup>6</sup> After finishing this article, we learned about a study by Juslin (1991), in which random samples were drawn from several natural environments. Overall, overconfidence in general knowledge was close to zero, consistent with this study.

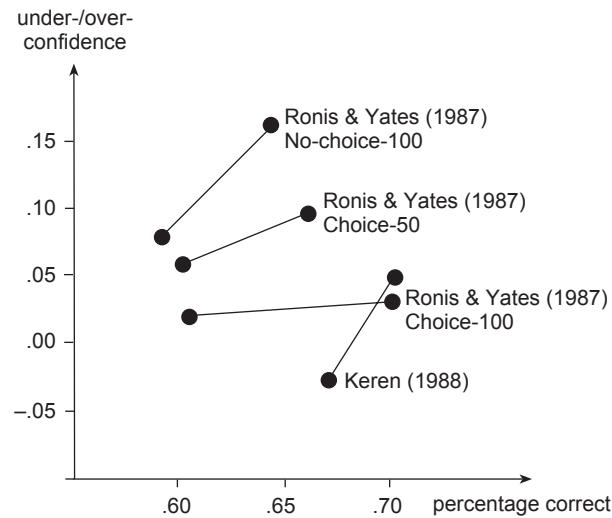


Figure 10. Reversal of the hard-easy effect in Ronis and Yates (1987) and Keren (1988).

an answer based on their knowledge, then (b) they selectively search for confirming memory (or for evidence disconfirming the alternative not chosen), and (c) this confirming evidence generates overconfidence. Between the subjects' choice of an answer and their confidence judgment, the authors asked the subjects to give reasons for the alternative chosen. Three groups of subjects were asked to write down one confirming reason, one disconfirming reason, or one of each, respectively. Reasons were given for half of the general-knowledge questions; otherwise, no reasons were given (control condition). If the confirmation bias explanation is correct, then asking for a contradicting reason (or both reasons) should decrease overconfidence and improve calibration. Asking for a confirming reason, however, should make no difference "since those instructions roughly simulate what people normally do" (Koriat et al., 1980, p. 111).

What does PMM theory predict? According to PMM theory, choice and confidence are inferred from the same activated cue. This cue is by definition a confirming reason. Therefore, the confirming-reason and the no-reason (control) tasks engage the same cognitive processes. The difference is only that in the former the supporting reason is written down. Similarly, the disconfirming-reason and both-reason tasks involve the same cognitive processes. Furthermore, PMM theory implies that there is no difference between the two pairs of tasks.

This result is shown in Table 4. In the first row we have the no-reason and confirming-reason tasks, which are equivalent. Here, only one cue is activated, which is confirming. There is no disconfirming cue. Now consider the second row, the disconfirming-reason and both-reason tasks, which are again equivalent. Both tasks are solved if one additional cue, which is disconfirming, can be activated. Thus, for PMM theory, the cue generation and testing cycle is started again, and cues are generated according to the hierarchy of cue validities and tested whether they can be activated for the problem at hand. The point is that the next cue that can be activated may turn out to be either confirming or disconfirming.

For simplicity, assume that the probability that the next activated cue turns out to be confirming or disconfirming is the same. If it is disconfirming, the cycle is stopped, and two cues in total have been activated, one confirming and one disconfirming. This stopping happens

Table 4  
*Predictions of PMM Theory for the Effects of Asking for a Disconfirming Reason*

Task	No. of cues activated	Cues activated			Predicted change in confidence	
		CON	DIS	Probability		
No; CON	1	CON	1	0	–	
DIS; both	2	DIS / CON	1	1	.5	Decrease
DIS; both	3	DIS / CON / CON	2	1	.25	Increase
DIS; both	4	DIS / CON / CON / CON	3	1	.125	Increase
DIS; both	>4	...	>3	1	.125	Increase

*Note.* CON = confirming reason; DIS = disconfirming reason; No = no reason; both = both reasons.

with probability .5, and it decreases both confidence and overconfidence. (Because the second cue activated has a smaller cue validity, however, confidence is not decreased below 50%.) If the second cue activated is again confirming, a third has to be activated, and the cue generation and testing cycle is entered again. If the third cue is disconfirming, the cycle stops with two confirming cues and one disconfirming cue activated, as shown in the third row of Table 4. This stopping is to be expected with probability .25. Because the second cue has higher cue validity than the third, disconfirming, cue, overall an increase in confidence and overconfidence is to be expected. If the third cue is again confirming, the same procedure is repeated. Here and in all subsequent cases confidence will increase. As shown in Table 4, the probabilities of an increase sum up to .5 (.25 + .125 + .125), which is the same as the probability of a decrease.

Thus, PMM theory leads to the prediction that, overall, asking for a disconfirming reason will not change confidence or overconfidence. As just shown, the confirmation-bias hypothesis, in contrast, predicts that asking for a disconfirming reason should decrease confidence and overconfidence.

What were the results of the Koriat study? In both crucial conditions, disconfirming reason and both reasons, the authors found only small and non-significant decreases of overconfidence (2% and 1%, respectively) and similar small improvements in calibration (.006 each, significant only in the disconfirming-reason task). These largely insignificant differences are consistent with the prediction by PMM theory that asking for a disconfirming reason makes no difference and are inconsistent with the confirmation-bias explanation. Further evidence comes from a replication of the Koriat study by Fischhoff and MacGregor (1982), who reported zero effects of disconfirming reasons.

To summarize, the effects on confidence of giving confirming and disconfirming reasons in the Koriat study can be both explained by and integrated into PMM theory. There is no need to postulate a confirmation bias.

*Dawes (1980).* Overconfidence has been attributed to people's tendency to "overestimate the power of our 'intellect' as opposed to that of our coding abilities." Such overestimation "has been reinforced by our realization that we have developed a technology capable of destroying ourselves" (Dawes 1980, p. 328). Dawes (1980) proposed that overconfidence is characteristic for general-knowledge questions but absent in perceptual tasks; he designed a series of experiments to test this proposal. PMM theory, however, gives no special treatment to perceptual tasks. On the contrary, it predicts overconfidence if perceptual tasks are selected for perceptual illusions—that is, for being

misleading—whereas zero overconfidence is to be expected if tasks are not selected. Pictures in textbooks on visual illusions are probably the set of items that produces the most extreme overconfidence yet demonstrated. Nevertheless, in a natural environment, perception is generally reliable.

Dawes reported inconsistent results. When perceptual stimuli were systematically constructed from a Square  $\times$  Circle matrix, as in the area task, and no selection for stimuli that generated perceptual illusions took place, overconfidence was close to zero (perception of areas of squares is quite well adapted in adults; see Gigerenzer & Richter, 1990). This result is predicted by both accounts. The anomaly arises with the second perceptual task used—judging which of two subsequent tones is longer.<sup>7</sup> If the second tone was longer, Dawes reported almost perfect calibration, but if the first tone was longer, subjects exhibited large overconfidence.

PMM theory predicts that in the inconsistent acoustic task, perceptual stimuli have been selected (albeit unwittingly) for a perceptual illusion. This is in fact the case. From the literature on time perception, we know that of two subsequently presented tones, the tone more recently heard appears to be longer. This perceptual illusion is known as the *negative presentation effect* (e.g., Fraisse, 1964; Sivyer & Finlay, 1982). It implies a smaller percentage of correct answers in the condition where the tone presented first was longer, because this tone is perceived to be shorter. A decrease in percentage correct in turn increases overconfidence. In Dawes' (1980) experiments, this is exactly the inconsistent condition where overconfidence occurred. Thus, from the perspective we propose, this inconsistent result can be reconciled.

*Keren (1988)*. A strict distinction between perceptual judgment and intellectual judgment cannot be derived from many views of perception, such as signal-detection theory (Tanner & Swets, 1954). Reflecting on this fact, Keren (1988) proposed a slightly modified hypothesis: The more perception-like a task is, the less overconfident and the better calibrated subjects will be. "As a task requires additional higher processing and transformation of the original sensory input, different kinds of possible cognitive distortions may exist (such as inappropriate inferences) that may limit the ability to accurately monitor our higher cognitive processes." (Keren, 1988, p. 99)

Keren (1988, Experiment 1) used general-knowledge questions and two kinds of perceptual tasks, one of them more difficult than the general-knowledge task, the other less difficult. Keren tested the hypothesis that confidence judgments in perceptual tasks are better calibrated than in general-knowledge tasks. He could not support it, however. Instead, he found an anomaly: The comparison between the general-knowledge task and the more difficult perceptual task reversed the hard-easy effect (see Figure 10). As derived in Prediction 5, this puzzling reversal is implied by PMM theory if the Landolt rings used in the more difficult perceptual task were not selected for perceptual illusions, as seems to be the case (Keren, 1988, p. 100).

Note that the kind of general-knowledge questions used (population of cities or countries, and distances between cities) would easily permit defining a reference class in a known environment and obtaining representative samples. But no representative sample of general-knowledge questions was generated. This lack makes the other predictions from PMM theory coincide with Keren's (1988): overconfidence in general-knowledge, and zero overconfidence in the two perceptual tasks. Results show this outcome, except for the large-gap Landolt rings condition, which generated considerable underconfidence. PMM theory cannot account for the latter, nor can the notion of degree of perception-likeness.

<sup>7</sup> Dawes' eye-color task is not dealt with here because it is a memory task, not a perceptual task.

A second perceptual task was letter identification. In Experiment 3, Keren (1988) used two letter-identification tasks, which were identical except that the exposure time of the letters to be recognized was either short or long. Mean percentages correct were 63.5 for short and 77.2 for long exposures. According to earlier explanations such as subjects' insensitivity to task difficulty, a hard-easy effect should result. According to PMM theory, however, the hard-easy effect should be zero, because both tasks were generated by the same sampling process (Prediction 4). In fact, Keren (1988, p. 112) reported that in both tasks, overconfidence was not significantly different from zero. He seems to have found no explanation for this disappearance of the hard-easy effect in a situation where differences in percentage correct were large.<sup>8</sup>

### *Mental Models as Probabilistic Syllogisms*

To the best of our knowledge, there is only one other mental models approach to confidence. May (1986, 1987) emphasized the role of mental models to understand the mechanism of confidence judgments and the role of misleading questions to cause overconfidence. Consider again the following question: "Which city has more inhabitants? (a) Hyderabad or (b) Islamabad." May proposed that subjects answer this question by constructing a mental model that can be expressed as a "probabilistic syllogistic inference" (May, 1986, p. 21):

"Most capitals have quite a lot of inhabitants.  
Islamabad is a capital.  
-----  
Presumably, Islamabad has quite a lot of inhabitants."

Replacing "most" by the probability  $p(\text{large}|\text{capital}) = 1 - \alpha$ , that is, the probability that a city has a large population if it is a capital, she made the following argument:

"The subjective probability therefore has to be '1 - alpha.' Given that the perception of that percentage is correct, the observed frequency of correct answers will be 1 - alpha. Even if there was random fluctuation of subjective probability and task difficulty, in the long run calibration is expected." (May, 1986, p. 20)

May (1986, 1987) did highly interesting analyses of individual general-knowledge questions—among others, direct tests of the use of the capital cue and the familiarity cue in questions of the

<sup>8</sup> Keren (1987, 1988; Wagenaar & Keren, 1985) also distinguished tasks in which the items are related (e.g., repeated weather forecasting) versus unrelated (e.g., typical general-knowledge questions). A similar distinction was made by Ronis and Yates (1987). PMM theory can connect Keren's distinction between two kinds of tasks with a model of cognitive processes involved in different tasks. In a set of unrelated items, a new PMM has to be constructed for each new item that cannot be answered by a local MM. This new PMM includes a new reference class, new target variable, and new cues and cue validities. This holds for reasoning about a set of typical general-knowledge questions. In contrast, the representative set of city questions used in our experiments implies that the PMMs for subsequent items include the same reference class, same target value, and same hierarchy of cues and cue validities but that different cues will be activated in different questions. Thus, in this framework, the distinction between related and unrelated items is neither a dichotomy nor a single continuum, but multidimensional. In general, a set of items can cue a series of PMMs that have (a) the same-different reference class, (b) the same-different target variable, (c) the same-different set of cues and cue validities, and (d) the same-different activated cues. Thus, at the other extreme of the typical general-knowledge task, there is a series of tasks that implies the construction of a succession of PMMs that are identical with respect to all four dimensions. An example is the repeated judgment of the frequency of correct answers in our experiments.

Hyderabad-Islamabad type. Both May and PMM theory share an emphasis on the mental models subjects use to construe their tasks. We discuss here two issues where we believe that May's position could be strengthened: the kind of mental model she proposed and the role of misleading questions.

First, we show that the probabilistic syllogism does not work in the sense she specified, that is, generating long-run calibration. We propose a working version. The general reason why the syllogism mental model does not work in a two-alternative task is that it does not deal with information about the alternative, that is, whether Hyderabad is known as a capital or noncapital or whether its status is unknown. Specifically, assume that the subject knows nothing about Hyderabad, as May did (1986, p. 19). The syllogism produces the choice  $a$  and the confidence judgment  $p(a = \text{large} | a = \text{capital})$ , where  $a$  stands for Islamabad. However, this confidence judgment is not equal to the long-run frequency of correct answers. The long-run frequency of correct answers  $p(a \text{ larger than } b | a = \text{capital})$  depends on both  $p(a = \text{large} | a = \text{capital})$  and  $p(b = \text{large} | b = \text{city})$ , where  $b$  stands for Hyderabad and "city" means no knowledge of whether  $b$  is a capital or a noncapital. For instance, the larger  $p(b = \text{large} | b = \text{city})$ , the smaller the long-run frequency of correct answers.

A mental model that generates confidence judgments that are well calibrated with long-run frequencies of correct answers can be constructed in at least two ways. First, the probabilistic syllogism can be supplanted by a second syllogism that uses our knowledge about the alternative—capital, noncapital, or just city (no knowledge of whether it is a capital or noncapital). Here is a numerical illustration of what can be called the *double-syllogism model*:

80% of capitals have a large population.  
Islamabad is a capital.

---

The probability is .80 that Islamabad has a large population.

40% of cities have a large population.  
Hyderabad is a city.

---

The probability is .40 that Hyderabad has a large population.

What is the long-run frequency of correct answers? There are four possible classes of events:  $a = \text{large}$  and  $b = \text{not large}$ ;  $b = \text{large}$  and  $a = \text{not large}$ ;  $a = \text{large}$  and  $b = \text{large}$ ; and  $a = \text{not large}$  and  $b = \text{not large}$ . Given the choice  $a$ , the first class of events signifies correct choices; the second, incorrect choices; and the third and fourth do not contain discriminating information. The long-run frequency of correct answers consists of the first class of events and of half of the third and fourth—on the assumption that half of the nondiscriminating cases will be correct and the other half incorrect. Thus, the long-run frequency of correct answers is  $p(a = \text{large}, b = \text{not large} | a = \text{capital}) + 1/2(p[a = \text{large}, b = \text{large} | a = \text{capital}] + p[a = \text{not large}, b = \text{not large} | a = \text{capital}])$ . We denote the probabilities from the first and the second syllogism as  $\alpha$  and  $\beta$ , respectively. Then, the long-run frequency of correct answers is  $\alpha(1 - \beta) + 1/2(\alpha\beta + [1 - \alpha][1 - \beta])$ , which is  $1/2(\alpha - \beta + 1)$ . For instance, if  $\alpha = \beta$  this probability is .50. Therefore, in the above double-syllogism model, the (calibrated) confidence that  $a$  is correct is  $1/2(\alpha - \beta + 1)$ . May (1986), in contrast, proposed  $\alpha$ . For the double syllogism, we get  $1/2(.80 - .40 + 1) = .70$ .

A second solution would be to dispense with the dichotomy of large versus small population and to use the cue validities as defined in PMM theory. Both changes would make May's mental models work and would lead to a mechanism that differs in an interesting way from that illus-

trated in Figure 3. In contrast to what we have proposed, May (1986) assumed that cues are activated even if knowledge from memory can be retrieved only for one alternative.

May (1986, 1987) also proposed that overconfidence is due to misleading items. The Islamabad-Hyderabad question is one example of a misleading item with less than 50% correct answers. Most subjects chose Islamabad, whereas Hyderabad has a much larger population. An extreme example is “Three fourths of the world’s cacao comes from (a) Africa or (b) South America” for which Fischhoff et al. (1977) reported only 4.8% correct answers. Fischhoff et al. showed that extreme overconfidence, such as odds greater than 50:1, are prevalent in what they called *deceptive* items (more than 73%) but still exist in nondeceptive items (less than 9%). Because of the latter finding, among other grounds, they concluded that misleading items “are not responsible for the extreme overconfidence effect” (p. 561). In contrast, May (1986, 1987) seems to have held that if misleading items were eliminated, overconfidence would be, too. But she also emphasized, as we do, the role of representative sampling.

We propose that the issue of misleading questions can be fully reduced to the issue of representative sampling from a reference class, which provides a deeper understanding of confidence in knowledge. As we have pointed out before, the same set of general-knowledge questions can be nonrepresentative with respect to one reference class (e.g., all cities in Germany), but representative with respect to a different reference class (e.g., sets of typical general-knowledge questions). The notion of misleading items does not capture this distinction, which is both essential to PMM theory as well as to an explanation for when and why overconfident subjects can quite realistically estimate their true relative frequencies of correct responses.

### *The Brunswikian Perspective*

PMM theory draws heavily on the Brunswikian notions of a natural environment known to an individual, reference classes in this environment, and representative sampling from a reference class. We went beyond the Brunswikian focus on achievement (rather than process) by providing a theoretical framework of the processes that determine choice, confidence, and frequency judgment.

Choice and confidence are a result of a cue-testing and activation cycle, which is analogous to Newell and Simon’s (1972) postulate that “problem solving takes place by search in a problem space” (p. 809). Furthermore, the emphasis on the structure of the task in PMM theory is similar to Newell and Simon’s proposition that “the structure of the task determines the possible structures of the problem space” (p. 789). Unlike PMM theorists, however, Newell and Simon also assumed in the tasks they studied (cryptarithmic, logic, and chess) a relatively simple mapping between the external structure of the task and the internal representation in a problem space (see Allport, 1975). Although it is cued by the task structure, we assume that a PMM (the functional equivalent of a problem space) has a large surplus structure (the reference class and the cues), which is taken from a known structure in the problem solver’s natural environment. The emphasis on the structure of everyday knowledge or environment (as distinguished from the task environment) has been most forcefully defended by Brunswik (see Gigerenzer, 1987). Although Newell and Simon (1972, p. 874) called Brunswik and Tolman “the real fore runners” of their work, they seem not to distinguish clearly between the notions of a probabilistic everyday environment and a task environment. This theory is an attempt to combine both views. Brunswik’s focus on achievement (during his behavioristic phase; see Leary, 1987) corresponds more closely to the part of research on probabilistic judgment that focuses on calibration, rather than on the underlying cognitive processes.

The importance of the cognitive representation of the task was studied by the Würzburg school and emphasized in Gestalt theoretical accounts of thinking (e.g., Duncker, 1935/1945), and this issue has recently regained favor (e.g., Brehmer, 1988; Hammond, Stewart, Brehmer, & Steinman, 1975). In their review, Einhorn and Hogarth (1981) emphasized that “the cognitive approach has been concerned primarily with *how* tasks are represented. The issue of *why* tasks are represented in particular ways has not yet been addressed.” (p. 57) PMM theory addresses this issue. Different tasks, such as confidence and frequency tasks, cue different reference classes and different probability cues from known environments. It is these environments that provide the particular representation, the PMM, of a task.

Many parts of PMM theory need further expansion, development, and testing. Open issues include the following: (a) What reference class is activated? For city comparisons, this question has a relatively clear answer, but in general, more than one reference class can be constructed to solve a problem. (b) Are cues always generated according to their rank in the cue validity hierarchy? Alternative models of cue generation could relax this strong assumption, assuming, for instance, that the first cue generated is the cue activated in the last problem. The latter would, however, decrease the percentage of correct answers. (c) What are the conditions under which we may expect PMMs to be well adapted? There exists a large body of neo-Brunswikian research that, in general, indicates good adaptation but also points out exceptions (e.g., K. Armelius, 1979; Brehmer & Joyce, 1988; Björkman, 1987; Hammond & Wascoe, 1980). (d) What are the conditions under which cue substitution without cue integration is superior to multiple cue integration? PMM theory assumes a pure cue substitution model—a cue that cannot be activated can be replaced by any other cue—without integration of two or more cues. We focused on the substitution and not the integration aspect of Brunswik’s vicarious functioning (see Gigerenzer & Murray, 1987, pp. 66–81), in contrast to the multiple regression metaphor of judgment. Despite its simplicity, the substitution model produces zero overconfidence and a large number of correct answers, if the PMM is well adapted. There may be more reasons for simple substitution models. Armelius and Armelius (1974), for instance, reported that subjects were well able to use ecological validities, but not the correlations between cues. If the latter is the case, then multiple cue integration may not work well.

We briefly indicate here that features emphasized by PMM theory, such as representative sampling and the confidence-frequency distinction, can also be crucial for probabilistic reasoning in other tasks.

In several Bayesian-type studies of revision of belief, representative (random) sampling from a reference class is a crucial issue. For instance, Gigerenzer, Hell, and Blank (1988) showed that subjects’ neglect of base rates in Kahneman and Tversky’s (1973) engineer-lawyer problem disappeared if subjects could randomly draw the descriptions from an urn. Similar results showing people’s sensitivity to the issue of representative versus selected sampling have been reported by Cosmides and Tooby (1990), Ginossar and Trope (1987), Grether (1980), Hansen and Donoghue (1977), and Wells and Harvey (1977), but see Nisbett and Borgida (1975).

This study has also demonstrated that judgments of single events can systematically differ from judgments of relative frequencies. Similar differences were found for other kinds of probabilistic reasoning (Gigerenzer, 1991a, 1991b). For instance, the “conjunction fallacy” has been established by asking subjects the probabilities of single events, such as whether “Linda” is more likely to be (a) a bank teller or (b) a bank teller and active in the feminist movement. Most subjects chose the latter, because the description of Linda was constructed to be representative of an active feminist. This judgment was called a conjunction fallacy because the probability of a con-

junction (bank teller and feminist) is never larger than the probability of one of its constituents. As in the engineer-lawyer problem, the representativeness heuristic was proposed to explain the “fallacy.” Fiedler (1988) and Tversky and Kahneman (1983), however, showed that the conjunction fallacy largely disappeared if people were asked for frequencies (e.g., “There are 100 persons like Linda. How many of them are ...?”) rather than probabilities of single events. Cosmides and Tooby (1990) showed a similar striking difference for people’s reasoning in a medical probability revision problem. The subjects’ task was to estimate the probability that people have a disease, given a positive test result, the base rate of the disease, the false-alarm rate, and the hit rate of the test. Originally, Casscells, Schoenberger, and Grayboys (1978) reported only 18% Bayesian answers when Harvard medical students and staff were asked for a single-event probability. (What is the probability that a person found to have a positive result actually has the disease?) When Cosmides and Tooby changed the task into a frequency task (How many people who test positive will actually have the disease?), 76% of subjects responded with the Bayesian answer. These results suggest that the mental models subjects construe to solve these reasoning problems were highly responsive to information crucial for probability and statistics—random versus selected sampling and single events versus frequencies in the long run.

### *Is Overconfidence a Bias According to Probability Theory?*

Throughout this article, we have avoided classifying judgments as either rational or biased, but instead focused on the underlying cognitive processes and how these explain extant data. Overconfidence is, however, usually classified as a bias and dealt with in chapters on “cognitive illusions” (e.g., Edwards & von Winterfeldt, 1986). Is overconfidence a bias according to probability theory?

Mathematical probability emerged around 1660 as a Janus-faced concept with three interpretations: observed frequencies of events, equal possibilities based on physical symmetry, and degrees of subjective certainty or belief. Frequencies originally came from mortality and natality data, sets of equiprobable outcomes from gambling, and the epistemic sense of belief proportioned to evidence from courtroom practices (Daston, 1988). Eighteenth-century mathematicians used “probability” in all three senses, whereas latter-day probabilists drew a bold line between the first two “objective” senses and the third “subjective” one. Today, mathematicians, statisticians, and philosophers are still wrangling over the proper interpretation of probability: Does it mean a relative frequency, a propensity, a degree of belief, a degree of evidentiary confirmation, or yet something else? Prominent thinkers can still be found in every camp, and it would be bold unto foolhardy to claim that any interpretation had a monopoly on reasonableness (Gigerenzer et al., 1989).

Overconfidence is defined as the difference between degrees of belief (subjective probabilities) and a relative frequency (percentage correct). Is a deviation between the probability that a particular answer is correct and the relative frequency of correct answers a bias or error, according to probability theory?

From the point of view of dedicated frequentists such as von Mises (1928/1957) and Neyman (1977), it is not. According to the frequentist interpretation (which is the dominant interpretation in statistics departments today), probability theory is about relative frequencies in the long run; it does not deal with degrees of beliefs concerning single events. For instance, when speaking of “the probability of death”:

[One] must not think of an individual, but of a certain class as a whole, e.g., “all insured men forty-one years old living in a given country and not engaged in certain dangerous occupations”. ... The phrase “probability of death,” when it refers to a single person, has no meaning at all for us. (von Mises, 1928/1957, p. 11)

For a frequentist, one cannot properly speak of a probability until a reference class has been defined. The statistician Barnard (1979), for instance, suggested that if one is concerned with the subjective probabilities of single events, such as confidence, one “should concentrate on the works of Freud and perhaps Jung rather than Fisher and Neyman” (p. 171). Thus, for frequentists, probability theory does not apply to single-event judgments like confidences, and therefore no statement about confidences can violate probability theory.

Moreover, even subjectivists would not generally think of a deviation between probabilities for single events and relative frequencies as a bias. The problem is whether and when a subjectivist, who rejects the identification of probability with objective frequency, should nonetheless make frequency the yardstick of good reasoning (for a discussion of conditions, see Kadane & Lichtenstein, 1982). The subjectivist Bruno de Finetti, for instance, emphatically stated in his early work that subjective probabilities of single events cannot be validated by objective probabilities:

However an individual evaluates the probability of a particular event, no experience can prove him right, or wrong; nor in general, could any conceivable criterion give any objective sense to the distinction one would like to draw, here, between right and wrong. (de Finetti, 1931/1989, p. 174)

We thus have to face a problem: Many cognitive psychologists think of overconfidence as a bias of reasoning, pointing to probability theory as justification. Many probabilists and statisticians, however, would reply that their interpretation of probability does not justify this label (see Hacking, 1965; Lad, 1984; Stegmüller, 1973).

PMM theory can offer a partial solution to this problem. First, it clarifies the distinction between confidence and frequency judgment and therewith directs attention to the comparison between estimated and true frequencies of correct answers. The latter avoids the previously stated problem. This comparison has not received much attention in research on confidence in knowledge. Second, PMM theory proposes a frequentist interpretation of degrees of belief: Both confidence and frequency judgments are based on memory about frequencies. Our view links both types of judgment but does not equate them. Rather, it specifies when to expect confidence and frequency judgments to diverge, and in what direction, and when they will converge. PMM theory integrates single-event probabilities into a frequentist framework: the Bayesian is Brunswikian.

### *Conclusions*

We conjecture that confidence in one’s knowledge of the kind studied here—immediate and spontaneous rather than a product of long-term reflection—is largely determined by the structure of the task and the structure of a corresponding, known environment in a person’s long-term memory. We provided experimental evidence for this hypothesis by showing how changes in the task (confidence vs. frequency judgment) and in the relationship between task and environment (selected vs. representative sampling) can make the two stable effects reported in the literature—overconfidence and the hard-easy effect—emerge, disappear, and invert at will. We have demonstrated a new phenomenon, the confidence-frequency effect. One cannot speak of a general overconfidence bias anymore, in the sense that it relates to deficient processes of cognition or mo-

tivation. In contrast, subjects seem to be able to make fine conceptual distinctions—confidence versus frequency—of the same kind as probabilists and statisticians do. Earlier attempts postulating general deficiencies in information processing or motivation cannot account for the experimental results predicted by PMM theory and confirmed in two experiments. PMM theory seems to be the first theory in this field that gives a coherent account of these various effects by focusing on the relation between the structure of the task, the structure of a corresponding environment, and a PMM.

## References

- Allport, D. A. (1975). The state of cognitive psychology. *Quarterly Journal of Experimental Psychology*, 27, 141–152.
- Allwood, C. M., & Montgomery, H. (1987). Response selection strategies and realism of confidence judgments. *Organizational Behavior and Human Decision Processes*, 39, 365–383.
- Anderson, N. H. (1986). A cognitive theory of judgment and decision. In B. Brehmer, H. Jungermann, P. Lourens, & G. Sevón (Eds.), *New directions in research on decision making* (pp. 63–108). Amsterdam: North-Holland.
- Angele, U., Beer-Binder, B., Berger, R., Bussmann, C., Kleinbölting, H., & Mansard, B. (1982). *Über- und Unterschätzung des eigenen Wissens in Abhängigkeit von Geschlecht und Bildungsstand*. [Overestimation and underestimation of one's knowledge as a function of sex and education]. Unpublished manuscript, University of Konstanz, Federal Republic of Germany.
- Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes*, 39, 133–144.
- Arkes, H. R., & Hammond, K. R. (Eds.). (1986). *Judgment and decision making: An interdisciplinary reader*. Cambridge, England: Cambridge University Press.
- Armelius, B., & Armelius, K. (1974). The use of redundancy in multiple-cue judgments: Data from a suppressor-variable-task. *American Journal of Psychology*, 87, 385–392.
- Armelius, K. (1979). Task predictability and performance as determinants of confidence in multiple-cue judgments. *Scandinavian Journal of Psychology*, 20, 19–25.
- Barnard, G. A. (1979). Discussion of the paper by Professors Lindley and Tversky and Dr. Brown. *Journal of the Royal Statistical Society of London*, 142 (Series A), 171–172.
- Björkman, M. (1987). A note on cue probability learning: What conditioning data reveal about cue contrast. *Scandinavian Journal of Psychology*, 28, 226–232.
- Brehmer, B. (1988). The development of social judgment theory. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 13–40). Amsterdam: North-Holland.
- Brehmer, B., & Joyce, C. R. B. (Eds.). (1988). *Human judgment: The SJT view*. Amsterdam: North-Holland.
- Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, 50, 255–272.
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62, 193–217.
- Brunswik, E. (1964). Scope and aspects of the cognitive problem. In *Contemporary approaches to cognition* (pp. 4–31). Cambridge, MA: Harvard University Press.
- Casscells, W., Schoenberger, A., & Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, 299, 999–1000.
- Cohen, L. J. (1989). *The philosophy of induction and probability*. Oxford, England: Clarendon Press.
- Cosmides, L., & Tooby, J. (1990, August). *Is the mind a frequentist?* Paper presented at the 31st Annual Meeting of the Psychonomics Society, New Orleans, LA.
- Daston, L. J. (1988). *Classical probability in the Enlightenment*. Princeton, NJ: Princeton University Press.
- Dawes, R. M. (1980). Confidence in intellectual judgments vs. confidence in perceptual judgments. In E. D. Lantermann & H. Feger (Eds.), *Similarity and choice: Papers in honor of Clyde Coombs* (pp. 327–345). Bern, Switzerland: Huber.
- de Finetti, B. (1989). Probabilism. *Erkenntnis*, 31, 169–223. (Original work published 1931)
- Duncker, K. (1945). On problem solving (L. S. Lees, Trans.). *Psychological Monographs*, 85 (5, Whole No. 270). (Original work published 1935)

- Edwards, W., & von Winterfeldt, D. (1986). On cognitive illusions and their implications. In H. R. Arkes & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (pp. 642–679). Cambridge, England: Cambridge University Press.
- Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology*, *32*, 53–88.
- Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, *50*, 123–129.
- Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge, England: Cambridge University Press.
- Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting*, *1*, 155–172.
- Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 552–564.
- Fraisse, P. (1964). *The psychology of time*. London: Eyre & Spottiswoode.
- Gigerenzer, G. (1984). External validity of laboratory experiments: The frequency-validity relationship. *American Journal of Psychology*, *97*, 185–195.
- Gigerenzer, G. (1987). Survival of the fittest probabilist: Brunswik, Thurstone, and the two disciplines of psychology. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution, Vol. 2: Ideas in the sciences*. Cambridge, MA: MIT Press.
- Gigerenzer, G. (1991a). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*, 254–267.
- Gigerenzer, G. (1991b). How to make cognitive illusions disappear: Beyond “heuristics and biases.” *European Review of Social Psychology*, *2*, 83–115.
- Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 513–525.
- Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.
- Gigerenzer, G., & Richter, H. R. (1990). Context effects and their interaction with development: Area judgments. *Cognitive Development*, *5*, 235–264.
- Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L. J., Beatty, J., & Krüger, L. (1989). *The empire of chance. How probability changed science and everyday life*. Cambridge, England: Cambridge University Press.
- Ginossar, Z., & Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology*, *52*, 464–474.
- Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics*, *95*, 537–557.
- Hacking, I. (1965). *Logic of statistical inference*. Cambridge, England: Cambridge University Press.
- Hammond, K. R., Stewart, T. R., Brehmer, B., & Steinmann, D. O. (1975). Social judgment theory. In M. F. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes*. San Diego, CA: Academic Press.
- Hammond, K. R., & Wascoe, N. E. (Eds.). (1980). Realizations of Brunswik’s representative design. *New Directions for the Methodology of Social and Behavioral Science*, *3*, 271–312.
- Hansen, R. D., & Donoghue, J. M. (1977). The power of consensus: Information derived from one’s own and others’ behavior. *Journal of Personality and Social Psychology*, *35*, 294–302.
- Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*, 107–112.
- Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, *108*, 356–388.
- Hintzman, D. L., Nozawa, G., & Irmscher, M. (1982). Frequency as a nonpropositional attribute of memory. *Journal of Verbal Learning and Verbal Behavior*, *21*, 127–141.
- Howell, W. C., & Burnett, S. (1978). Uncertainty measurement: A cognitive taxonomy. *Organizational Behavior and Human Performance*, *22*, 45–68.
- Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.
- Juslin, P. (1991). *Well-calibrated general knowledge: An ecological inductive approach to realism of confidence*. Manuscript submitted for publication, Uppsala, Sweden.
- Kadane, J. B., & Lichtenstein, S. (1982). *A subjectivist view of calibration* (Rep. No. 82–86). Eugene, OR: Decision Research.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.
- Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 493–508). Cambridge, England: Cambridge University Press.
- Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, *39*, 98–114.

- Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica*, *67*, 95–119.
- Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118.
- Kyburg, H. E. (1983). Rational belief. *Behavioral and Brain Sciences*, *6*, 231–273.
- Lad, F. (1984). The calibration question. *British Journal of the Philosophy of Science*, *35*, 213–221.
- Leary, D. E. (1987). From act psychology to probabilistic functionalism: The place of Egon Brunswik in the history of psychology. In M. G. Ash & W. R. Woodward (Eds.), *Psychology in twentieth-century thought and society* (pp. 115–142). Cambridge, England: Cambridge University Press.
- Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior and Human Performance*, *20*, 159–183.
- Lichtenstein, S., & Fischhoff, B. (1981). *The effects of gender and instruction on calibration* (Tech. Rep. No. PTR-1092-81-7). Eugene, OR: Decision Research.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: Cambridge University Press.
- MacGregor, D., & Slovic, P. (1986). Perceived acceptability of risk analysis as a decision-making approach. *Risk Analysis*, *6*, 245–256.
- May, R. S. (1986). Overconfidence as a result of incomplete and wrong knowledge. In R. W. Scholz (Ed.), *Current issues in West German decision research* (pp. 13–30). Frankfurt a.M., Germany: Lang.
- May, R. S. (1987). *Realismus von subjektiven Wahrscheinlichkeiten: Eine kognitionspsychologische Analyse inferentieller Prozesse beim Overconfidence-Phänomen* [Calibration of subjective probabilities: A cognitive analysis of inference processes in overconfidence]. Frankfurt a.M., Germany: Lang.
- Mayselless, O., & Kruglanski, A. W. (1987). What makes you so sure? Effects of epistemic motivations on judgmental confidence. *Organizational Behavior and Human Decision Processes*, *39*, 162–183.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*, 97–131.
- Nisbett, R. E., & Borgida, E. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology*, *32*, 932–943.
- Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, *72*, 407–418.
- Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, *40*, 193–218.
- Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.
- Schönemann, P. H. (1983). Some theory and results for metrics for bounded response scales. *Journal of Mathematical Psychology*, *27*, 311–324.
- Shafer, G. (1978). Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for the History of Exact Sciences*, *19*, 309–370.
- Sivyer, M., & Finlay, D. (1982). Perceived duration of auditory sequences. *Journal of General Psychology*, *107*, 209–217.
- Statistisches Bundesamt. (1986). *Statistisches Jahrbuch 1986 für die Bundesrepublik Deutschland* [Statistical yearbook 1986 for the Federal Republic of Germany]. Stuttgart, Germany: Kohlhammer.
- Stegmüller, W. (1973). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Bd. IV: Personelle und Statistische Wahrscheinlichkeit. Teil E* [Problems and results of philosophy of science and analytical philosophy. Vol. IV: Personal and statistical probability. Part E]. Berlin, Germany: Springer.
- Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*, 401–409.
- Teigen, K. H. (1974). Overestimation of subjective probabilities. *Scandinavian Journal of Psychology*, *15*, 56–62.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.
- von Mises, R. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*. Berlin, Germany: Springer. (Translated and reprinted as *Probability, statistics, and truth*. New York: Dover, 1957)
- von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge, England: Cambridge University Press.
- Wagenaar, W., & Keren, G. B. (1985). Calibration of probability assessments by professional blackjack dealers, statistical experts, and lay people. *Organizational Behavior and Human Decision Processes*, *36*, 406–416.
- Wells, G. L., & Harvey, J. H. (1977). Do people use consensus information in making causal attributions? *Journal of Personality and Social Psychology*, *35*, 279–293.

- Zacks, R. T., Hasher, L., & Sanft, H. (1982). Automatic encoding of event frequency: Further findings. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 8, 106–116.
- Zimmer, A. C. (1983). Verbal vs. numerical processing by subjective probabilities. In R. W. Scholz (Ed.), *Decision making under uncertainty* (pp. 159–182). Amsterdam: North-Holland.
- Zimmer, A. C. (1986). What uncertainty judgments can tell about the underlying subjective probabilities. *Uncertainty in Artificial Intelligence*, 249–258.