The Weighing of Evidence and the Determinants of Confidence

DALE GRIFFIN AND AMOS TVERSKY

The University of Waterloo and Stanford University

The pattern of overconfidence and underconfidence observed in studies of intuitive judgment is explained by the hypothesis that people focus on the strength or extremeness of the available evidence (e.g., the warmth of a letter or the size of an effect) with insufficient regard for its weight or credence (e.g., the credibility of the writer or the size of the sample). This mode of judgment yields overconfidence when strength is high and weight is low, and underconfidence when strength is low and weight is high. We first demonstrate this phenomenon in a chance setup where strength is defined by sample proportion and weight is defined by sample size, and then extend the analysis to more complex evidential problems, including general knowledge questions and predicting the behavior of self and of others. We propose that people's confidence is determined by the balance of arguments for and against the competing hypotheses, with insufficient regard for the weight of the evidence. We show that this account can explain the effect of item difficulty on overconfidence, and we relate the observed discrepancy between confidence judgments and frequency estimates to the illusion of validity. Finally, we contrast the present account with a frequentistic model of confidence proposed by Gigerenzer and his colleagues, and present data that refute their model. © 1992 Academic Press, Inc.

The weighing of evidence and the formation of belief are basic elements of human thought. The question of how to evaluate evidence and assess confidence has been addressed from a normative perspective by philosophers and statisticians; it has also been investigated experimentally by psychologists and decision researchers. One of the major findings that has emerged from this research is that people are often more confident in their judgments than is warranted by the facts. Overconfidence is not limited to lay judgment or laboratory experiments. The well-publicized observation that more than two-thirds of small businesses fail within 4 years (Dun & Bradstreet, 1967) suggests that many entrepreneurs overestimate their probability of success (Cooper, Woo, & Dunkelberg, 1988). With some notable exceptions, such as weather forecasters (Murphy & Winkler,

This work was supported by a NSERC research grant to the first author and by Grant 89-0064 from the Air Force Office of Scientific Research to the second author. The paper has benefited from discussions with Robyn Dawes, Baruch Fischhoff and Daniel Kahneman. Reprint requests should be addressed to Amos Tversky, Department of Psychology, Stanford University, Jordan Hall, Building 420, Stanford, CA 94305.

0010-0285/92 $9.00

1977) who receive immediate frequentistic feedback and produce realistic forecasts of precipitation, overconfidence has been observed in judgments of physicians (Lusted, 1977), clinical psychologists (Oskamp, 1965), lawyers (Wagenaar & Keren, 1986), negotiators (Neale & Bazerman, 1990), engineers (Kidd, 1970), and security analysts (Staël von Holstein, 1972). As one critic described expert prediction, "often wrong but rarely in doubt."

Overconfidence is common but not universal. Studies of calibration have found that with very easy items, overconfidence is eliminated, and underconfidence is often observed (Lichtenstein, Fischhoff, & Phillips, 1982). Furthermore, studies of sequential updating have shown that posterior probability estimates commonly exhibit conservatism or underconfidence (Edwards, 1968). In the present paper, we investigate the weighting of evidence and propose an account that explains the pattern of overconfidence and underconfidence observed in the literature.[1]

THE DETERMINANTS OF CONFIDENCE

The assessment of confidence or degree of belief in a given hypothesis typically requires the integration of different kinds of evidence. In many problems, it is possible to distinguish between the strength, or extremeness, of the evidence and its weight, or predictive validity. When we evaluate a letter of recommendation for a graduate student written by a former teacher, we may wish to consider two separate aspects of the evidence: (i) how positive or warm is the letter? and (ii) how credible or knowledgeable is the writer? The first question refers to the strength or extremeness of the evidence, whereas the second question refers to its weight or credence. Similarly, suppose we wish to evaluate the evidence for the hypothesis that a coin is biased in favor of heads rather than in favor of tails. In this case, the proportion of heads in a sample reflects the strength of evidence for the hypothesis in question, and the size of the sample reflects the credence of these data. The distinction between the strength of evidence and its weight is closely related to the distinction between the size of an effect (e.g., a difference between two means) and its reliability (e.g., the standard error of the difference). Although it is not always possible to decompose the impact of evidence into the separate contributions of strength and weight, there are many contexts in which they can be varied independently. A strong or a weak recommendation may come from a reliable or unreliable source, and the same proportion of heads can be observed in a small or large sample.

[1] A person is said to exhibit overconfidence if she overestimates the probability of her favored hypothesis. The appropriate probability estimate may be determined empirically (e.g., by a person's hit rate) or derived from an appropriate model.

Statistical theory and the calculus of chance prescribe rules for combining strength and weight. For example, probability theory specifies how sample proportion and sample size combine to determine posterior probability. The extensive experimental literature on judgment under uncertainty indicates that people do not combine strength and weight in accord with the rules of probability and statistics. Rather, intuitive judgments are overly influenced by the degree to which the available evidence is representative of the hypothesis in question (Dawes, 1988; Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980). If people were to rely on representativeness alone, their judgments (e.g., that a person being interviewed will be a successful manager) would depend only on the strength of their impression (e.g., the degree to which the individual in question "looks like" a successful manager) with no regard for other factors that control predictive validity. In many situations, however, it appears that people do not neglect these factors altogether. Instead, we propose, people focus on the strength of the evidence—as they perceive it—and then make some adjustment in response to its weight.

In evaluating a letter of recommendation, we suggest, people first attend to the warmth of the recommendation and then make allowance for the writer's limited knowledge. Similarly, when judging whether a coin is biased in favor of heads or in favor of tails, people focus on the proportion of heads in the sample and then adjust their judgment according to the number of tosses. Because such an adjustment is generally insufficient (Slovic & Lichtenstein, 1971; Tversky & Kahneman, 1974), the strength of the evidence tends to dominate its weight in comparison to an appropriate statistical model. Furthermore, the tendency to focus on the strength of the evidence leads people to underutilize other variables that control predictive validity, such as base rate and discriminability. This treatment combines judgment by representativeness, which is based entirely on the strength of an impression, with an anchoring and adjustment process that takes the weight of the evidence into account, albeit insufficiently. The role of anchoring in impression formation has been addressed by Quattrone (1982).

This hypothesis implies a distinctive pattern of overconfidence and underconfidence. If people are highly sensitive to variations in the extremeness of evidence and not sufficiently sensitive to variations in its credence or predictive validity, then judgments will be overconfident when strength is high and weight is low, and they will be underconfident when weight is high and strength is low. As is shown below, this hypothesis serves to organize and summarize much experimental evidence on judgment under uncertainty.

Consider the prediction of success in graduate school on the basis of a letter of recommendation. If people focus primarily on the warmth of the recommendation with insufficient regard for the credibility of the writer, or the correlation between the predictor and the criterion, they will be overconfident when they encounter a glowing letter based on casual contact, and they will be underconfident when they encounter a moderately positive letter from a highly knowledgeable source. Similarly, if people's judgments regarding the bias of a coin are determined primarily by the proportion of heads and tails in the sample with insufficient regard for sample size, then they will be overconfident when they observe an extreme proportion in a small sample, and underconfident when they observe a moderate proportion in a large sample.

In this article, we test the hypothesis that overconfidence occurs when strength is high and weight is low, and underconfidence occurs when weight is high and strength is low. The first three experiments are concerned with the evaluation of statistical hypotheses, where strength of evidence is defined by sample proportion. In the second part of the paper, we extend this hypothesis to more complex evidential problems and investigate its implications for judgments of confidence.

EVALUATING STATISTICAL HYPOTHESES

Study 1: Sample Size

We first investigate the relative impact of sample proportion (strength) and sample size (weight) in an experimental task involving the assessment of posterior probability. We presented 35 students with the following instructions:

> Imagine that you are spinning a coin, and recording how often the coin lands heads and how often the coin lands tails. Unlike tossing, which (on average) yields an equal number of heads and tails, spinning a coin leads to a bias favoring one side or the other because of slight imperfections on the rim of the coin (and an uneven distribution of mass). Now imagine that you know that this bias is 3/5. It tends to land on one side 3 out of 5 times. But you do not know if this bias is in favor of heads or in favor of tails.

Subjects were then given various samples of evidence differing in sample size (from 3 to 33) and in the number of heads (from 2 to 19). All samples contained a majority of heads, and subjects were asked to estimate the probability (from .5 to 1) that the bias favored heads (*H*) rather than tails (*T*). Subjects received all 12 combinations of sample proportion and sample size shown in Table 1. They were offered a prize of $20 for the person whose judgments most closely matched the correct values.

Table 1 also presents, for each sample of data (*D*), the posterior probability for hypothesis *H* (a 3:2 bias in favor of heads) computed according to Bayes' Rule. Assuming equal prior probabilities, Bayes' Rule yields

TABLE 1
Stimuli and Responses for Study 1

Number of heads (h)	Number of tails (t)	Sample size (n)	Posterior probability $P(H\|D)$	Median confidence (in %)
2	1	3	.60	63.0
3	0	3	.77	85.0
3	2	5	.60	60.0
4	1	5	.77	80.0
5	0	5	.88	92.5
5	4	9	.60	55.0
6	3	9	.77	66.9
7	2	9	.88	77.0
9	8	17	.60	54.5
10	7	17	.77	59.5
11	6	17	.88	64.5
19	14	33	.88	60.0

$$\log\left(\frac{P(H|D)}{P(T|D)}\right) = n\left(\frac{h-t}{n}\right)\log\left(\frac{.6}{.4}\right),$$

where h and t are the number of heads and tails, respectively, and $n = h + t$ denotes sample size. The first term on the right-hand side, n, represents the weight of evidence. The second term, the difference between the proportion of heads and tails in the sample, represents the strength of the evidence for H against T. The third term, which is held constant in this study, is the discriminability of the two hypotheses, corresponding to d' in signal detection theory. Plotting equal-support lines for strength and weight in logarithmic coordinates yields a family of parallel straight lines with a slope of -1, as illustrated by the dotted lines in Fig. 1. (To facilitate interpretation, the strength dimension is defined as h/n which is linearly related to $(h-t)/n$.) Each line connects all data sets that provide the same support for hypothesis H. For example, a sample of size 9 with 6 heads and 3 tails, and a sample of size 17 with 10 heads and 7 tails, yields the same posterior probability (.77) for H over T. Thus the point (9, 6/9) and the point (17, 10/17) both lie on the upper line. Similarly, the lower line connects the data sets that yield a posterior probability of .60 in favor of H (see Table 1).

To compare the observed judgments with Bayes' Rule, we first transformed each probability judgment into log odds and then, for each subject as well as the median data, regressed the logarithm of these values against the logarithms of strength, $(h-t)/n$, and of weight, n, separately for each subject. The regressions fit the data quite well: multiple R was .95 for the median data and .82 for the median subject. According to Bayes' Rule,

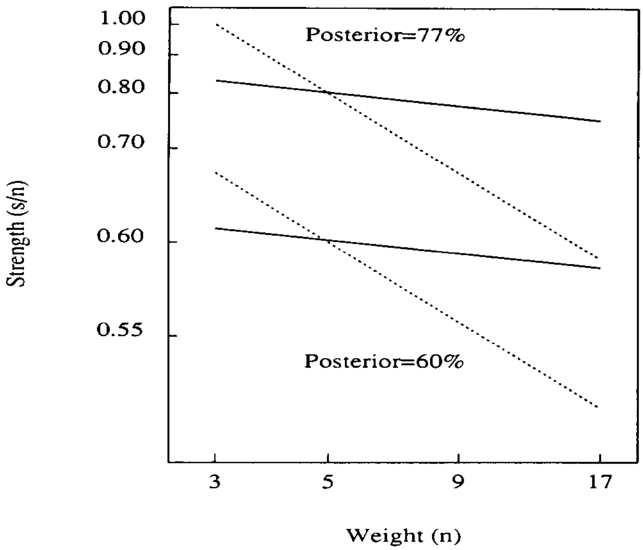

FIG. 1. Equal support lines for strength and sample size.

the regression weights for strength and weight in this metric are equal (see Fig. 1). In contrast, the regression coefficient for strength was larger than the regression coefficient for weight for 30 out of 35 subjects ($p < .001$ by sign test). Across subjects, the median ratio of these coefficients was 2.2 to 1 in favor of strength.[2] For the median data, the observed regression weight for strength (.81) was almost 3 times larger than that for weight (.31).

The equal-support lines obtained from the regression analysis are plotted in Fig. 1 as solid lines. The comparison of the two sets of lines highly reveal two noteworthy observations. First, the intuitive lines are much shallower than the Bayesian lines, indicating that the strength of evidence dominates its weight. Second, for a given level of support (e.g., 60% or 77%), the Bayesian and the intuitive lines cross, indicating overconfidence where strength is high and weight is low, and underconfidence where strength is low and weight is high. As is seen later, the crossing point is determined primarily by the discriminability of the competing hypotheses (d').

Figure 2 plots the median confidence for a given sample of evidence as

[2] To explore the effect of the correlation between strength and weight, we replicated our experiment with another set of stimuli that were selected to have a smaller correlation between the two independent variables ($r = -.27$ as compared to $r = -.64$). The results for this set of stimuli were remarkably similar to those reported in the text, i.e., the regression weights for the median data yielded a ratio of nearly 2 to 1 in favor of strength.

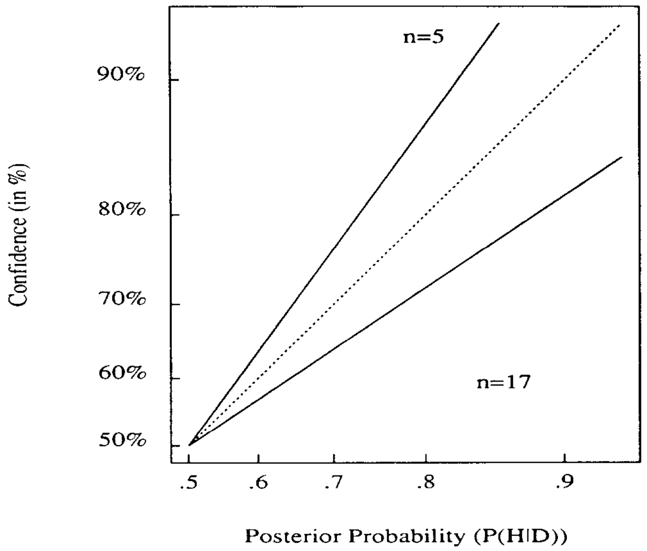

FIG. 2. Sample size and confidence.

a function of the (Bayesian) posterior probability for two separate sample sizes. The best-fitting lines were calculated using the log odds metric. If the subjects were Bayesian, the solid lines would coincide with the dotted line. Instead, intuitive judgments based on the small sample ($n = 5$) were overconfident, whereas the judgments based on the larger sample ($n = 17$) were underconfident.

The results described in Table 1 are in general agreement with previous results that document the non-normative nature of intuitive judgment (for reviews see, e.g., Kahneman, Slovic, & Tversky, 1982; von Winterfeldt & Edwards, 1986). Moreover, they help reconcile apparently inconsistent findings. Edwards and his colleagues (e.g., Edwards, 1968), who used a sequential updating paradigm, argued that people are conservative in the sense that they do not extract enough information from sample data. On the other hand, Tversky & Kahneman (1971), who investigated the role of sample size in researchers' confidence in the replicability of their results, concluded that people (even those trained in statistics) make radical inferences on the basis of small samples. Figures 1 and 2 suggest how the dominance of sample proportion over sample size could produce both findings. In some updating experiments conducted by Edwards, subjects were exposed to large samples of data typically of moderate strength. This is the context in which we expect underconfidence or conservatism. The situations studied by Tversky & Kahneman, on the other hand, involve moderately strong effects based on fairly small samples. This is the context in which overconfidence is likely to prevail. Both conservatism and overconfidence, therefore, can be generated by a common bias in the weighting of evidence, namely the dominance of strength over weight.

As was noted earlier, the tendency to focus on the strength of the evidence leads people to neglect or underweight other variables, such as the prior probability of the hypothesis in question or the discriminability of the competing hypotheses. These effects are demonstrated in the following two studies. All three studies reported in this section employ a within-subject design, in which both the strength of the evidence and the mitigating variable (e.g., sample size) are varied within subjects. This procedure may underestimate the dominance of strength because people tend to respond to whatever variable is manipulated within a study whether or not it is normative to do so (Fischhoff & Bar-Hillel, 1984). Indeed, the neglect of sample size and base-rate information has been most pronounced in between-subject comparisons (Kahneman & Tversky, 1972).

Study 2: Base Rate

Considerable research has demonstrated that people tend to neglect background data (e.g., base rates) in the presence of specific evidence (Kahneman, Slovic, & Tversky, 1982; Bar-Hillel, 1983). This neglect can lead either to underconfidence or overconfidence, as is shown below. We asked 40 students to imagine that they had three different foreign coins, each with a known bias of 3:2. As in Study 1, subjects did not know if the bias of each coin was in favor of heads (*H*) or in favor of tails (*T*). The subjects' prior probabilities of the two hypotheses (*H* and *T*) were varied. For one-half of the subjects, the probability of *H* was .50 for one type of coin, .67 for a second type of coin, and .90 for a third type of coin. For the other half of the subjects, the prior probabilities of *H* were .50, .33, and .10. Subjects were presented with samples of size 10, which included from 5 to 9 heads. They were then asked to give their confidence (in %) that the coin under consideration was biased in favor of heads. Again, a $20 prize was offered for the person whose judgments most closely matched the correct values. Table 2 summarizes the sample data, the posterior probability for each sample, and subjects' median confidence judgments. It is clear that our subjects overweighted strength of evidence and underweighted the prior probability.

Figure 3 plots median judgments of confidence as a function of (Bayesian) posterior probability for high (.90) and low (.10) prior probabilities of *H*. The figure also displays the best-fitting lines for each condition. It is evident from the figure that subjects were overconfident in the low base rate condition and underconfident in the high base rate condition.

These results are consistent with Grether's (1980; 1990) studies on the

TABLE 2
Stimuli and Responses for Study 2

Number of heads (out of 10)	Prior probability (Base rate)	Posterior probability $P(H\|D)$	Median confidence (in %)
5	9:1	.90	60.0
6	9:1	.95	70.0
7	9:1	.98	85.0
8	9:1	.99	92.5
9	9:1	.996	98.5
5	2:1	.67	55.0
6	2:1	.82	65.0
7	2:1	.91	71.0
8	2:1	.96	82.5
9	2:1	.98	90.0
5	1:1	.50	50.0
6	1:1	.69	60.0
7	1:1	.84	70.0
8	1:1	.92	80.0
9	1:1	.96	90.0
5	1:2	.33	33.0
6	1:2	.53	50.0
7	1:2	.72	57.0
8	1:2	.85	77.0
9	1:2	.93	90.0
5	1:9	.10	22.5
6	1:9	.20	45.0
7	1:9	.36	60.0
8	1:9	.55	80.0
9	1:9	.74	85.0

role of the representativeness heuristic in judgments of posterior probability. Unlike the present study, where both prior probabilities and data were presented in numerical form, Grether's procedure involved random sampling of numbered balls from a bingo cage. He found that subjects overweighted the likelihood ratio relative to prior probability, as implied by representativeness, and that monetary incentives reduced but did not eliminate base rate neglect. Grether's results, like those found by Camerer (1990) in his extensive study of market trading, contradict the claim of Gigerenzer, Hell, and Blank (1988) that explicit random sampling eliminates base rate neglect. Evidence that explicit random sampling alone does not reduce base rate neglect is presented in Griffin (1991).

Our analysis implies that people are prone to overconfidence when the base rate is low and to underconfidence when the base rate is high. Dun-

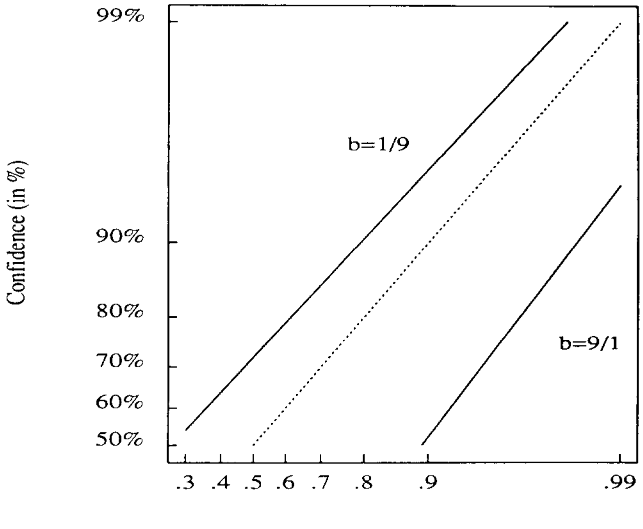

FIG. 3. Base rate and confidence.

ning, Griffin, Milojkovic, and Ross (1990) observed this pattern in a study of social prediction. In their study, each subject interviewed a target person before making predictions about the target's preferences and behavior (e.g., "If this person were offered a free subscription, which magazine would he choose: *Playboy* or *New York Review of Books?*"). The authors presented each subject with the empirically derived estimates of the base rate frequency of the responses in question (e.g., that 68% of prior respondents preferred *Playboy*). To investigate the effect of empirical base rates, Dunning et al. analyzed separately the predictions that agreed with the base rate (i.e., "high" base rate predictions) and the predictions that went against the base rate (i.e., "low" base rate predictions). Overconfidence was much more pronounced when base rates were low (confidence = 72%, accuracy = 49%) than when base rates were high (confidence = 79%, accuracy = 75%). Moreover, for items with base rates that exceeded 75%, subjects' predictions were actually underconfident. This is exactly the pattern implied by the hypothesis that subjects evaluate the probability that a given person would prefer *Playboy* over the *New York Review of Books* on the basis of their impression of that person with little or no regard for the empirical base rate, that is, the relative popularity of the two magazines in the target population.

Study 3: Discriminability

When we consider the question of which of two hypotheses is true, confidence should depend on the degree to which the data fit one hypothesis better than the other. However, people seem to focus on the strength of evidence for a given hypothesis and neglect how well the same evidence fits an alternate hypothesis. The Barnum effect is a case in point. It is easy to construct a personality sketch that will impress many people as a fairly accurate description of their own characteristics because they evaluate the description by the degree to which it fits their personality with little or no concern for whether it fits others just as well (Forer, 1949). To explore this effect in a chance setup, we presented 50 students with evidence about two types of foreign coins. Within each type of coin, the strength of evidence (sample proportion) varied from 7/12 heads to 10/12 heads. The two types of coins differed in their characteristic biases. Subjects were instructed:

> Imagine that you are spinning a foreign coin called a *quinta*. Suppose that half of the quintas (the "X" type) have a .6 bias towards Heads (that is, Heads comes up on 60% of the spins for X-quintas) and half of the quintas (the "Y" type) have a .75 bias toward Tails (that is, Tails comes up on 75% of the spins for Y-quintas). Your job is to determine if this is an X-quinta or a Y-quinta.

They then received the samples of evidence displayed in Table 3. After they gave their confidence that each sample came from an X-quinta or a Y-quinta, subjects were asked to make the same judgments for A-libnars (which have a .6 bias toward heads) and B-libnars (which have a .5 chance of heads). The order of presentation of coins was counterbalanced.

Table 3 summarizes the sample data, the posterior probability for each sample, and subjects' median confidence judgments. The comparison of the confidence judgments to the Bayesian posterior probabilities indicates that our subjects focused primarily on the degree to which the data fit the favored hypothesis with insufficient regard for how well they fit the alternate hypothesis (Fischhoff & Beyth-Marom, 1983). Figure 4 plots sub-

TABLE 3
Stimuli and Responses for Study 3

Number of heads (out of 12)	Separation of hypotheses (d')	Posterior probability $P(H\|D)$	Median confidence (in %)
7	.6 vs .5	.54	55.0
8	.6 vs .5	.64	66.0
9	.6 vs .5	.72	75.0
10	.6 vs .5	.80	85.0
7	.6 vs .25	.95	65.0
8	.6 vs .25	.99	70.0
9	.6 vs .25	.998	80.0
10	.6 vs .25	.999	90.0

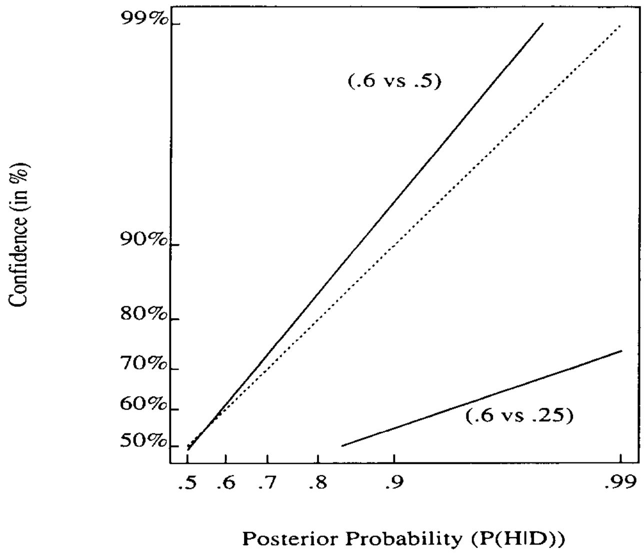

FIG. 4. Discriminability and confidence.

jects' median confidence judgments against the Bayesian posterior probability both for low discriminability and high discriminability comparisons. When the discriminability between the hypotheses was low (when the coin's bias was either .6 or .5) subjects were slightly overconfident, when the discriminability between the hypotheses was high (when the bias was either .6 or .25) subjects were grossly underconfident.

In the early experimental literature on judgments of posterior probability, most studies (e.g., Peterson, Schneider, & Miller, 1965) examined symmetric hypotheses that were highly discriminable (e.g., 3:2 versus 2:3) and found consistent underconfidence. In accord with our hypothesis, however, studies which included pairs of hypotheses of low discriminability found overconfidence. For example, Peterson and Miller (1965) found overconfidence in posterior probability judgments when the respective ratios were 3:2 and 3:4, and Phillips and Edwards (1966) found overconfidence when the ratios were 11:9 and 9:11.

CONFIDENCE IN KNOWLEDGE

The preceding section shows that people are more sensitive to the strength of evidence than to its weight. Consequently, people are overconfident when strength is high and weight is low, and underconfident when strength is low and weight is high. This conclusion, we propose, applies not only to judgments about chance processes such as coin spinning, but also to judgments about uncertain events such as who will win an upcoming election, or whether a given book will make the best-seller list. When people assess the probability of such events they evaluate, we suggest, their impression of the candidate or the book. These impressions may be based on a casual observation or on extensive knowledge of the preferences of voters and readers. In an analogy to a chance setup, the extremeness of an impression may be compared to sample proportion, and the credence of an impression may correspond to the size of the sample, or to the discriminability of the competing hypotheses. If people focus on the strength of the impression with insufficient appreciation of its weight, then the pattern of overconfidence and underconfidence observed in the evaluation of chance processes should also be present in evaluations of non-statistical evidence.

In this section, we extend this hypothesis to complex evidential problems where strength and weight cannot be readily defined. We first compare the prediction of self and of others. Next, we show how the present account gives rise to the "difficulty effect." Finally, we explore the determinants of confidence in general-knowledge questions, and relate the confidence-frequency discrepancy to the illusion of validity.

Study 4: Self versus Other

In this study, we ask people to predict their own behavior, about which they presumably know a great deal, and the behavior of others, about which they know less. If people base their confidence primarily on the strength of their impression with insufficient regard for its weight, we expect more overconfidence in the prediction of others than in the prediction of self.

Fourteen pairs of same-sex students, who did not know each other, were asked to predict each other's behavior in a task involving risk. They were first given 5 min to interview each other, and then they sat at individual computer terminals where they predicted their own and their partner's behavior in a Prisoner's Dilemma-type game called "The Corporate Jungle." On each trial, participants had the option of "merging" their company with their partner's company (i.e, cooperating), or "taking over" their partner's company (i.e., competing). If one partner tried to merge and the other tried to take over, the cooperative merger took a steep loss and the corporate raider made a substantial gain. However, if both partners tried a takeover on the same trial, they both suffered a loss. There were 20 payoff matrices, some designed to encourage cooperation and some designed to encourage competition.

Subjects were asked to predict their own behavior for 10 of the payoff matrices and the behavior of the person they had interviewed for the other 10. The order of the two tasks was counterbalanced, and each payoff matrix appeared an equal number of times in each task. In addition to predicting cooperation or competition for each matrix, subjects indicated their confidence in each prediction (on a scale from 50% to 100%). Shortly after the completion of the prediction task, subjects played 20 trials against their opponents, without feedback, and received payment according to the outcomes of the 20 trials.

The analysis is based on 25 subjects who completed the entire task. Overall, subjects were almost equally confident in their self predictions (M = 84%) and in their predictions of others (M = 83%), but they were considerably more accurate in predicting their own behavior (M = 81%) than in predicting the behavior of others (M = 68%). Thus, people exhibited considerable overconfidence in predictions of others, but were relatively well-calibrated in predicting themselves (see Fig. 5).

In some circumstances, where the strength of evidence is not extreme, the prediction of one's own behavior may be underconfident. In the case of a job choice, for example, underconfidence may arise if a person has good reasons for taking job A and good reasons for taking job B, but fails to appreciate that even a small advantage for job A over B would generally lead to the choice of A. If confidence in the choice of A over B reflects the balance of arguments for the two positions (Koriat, Lichtenstein, & Fischhoff, 1980), then a balance of 2 to 1 would produce confidence of about 2/3, although the probability of choosing A over B is likely to be higher. Over the past few years, we have discreetly approached colleagues faced with a choice between job offers, and asked them to

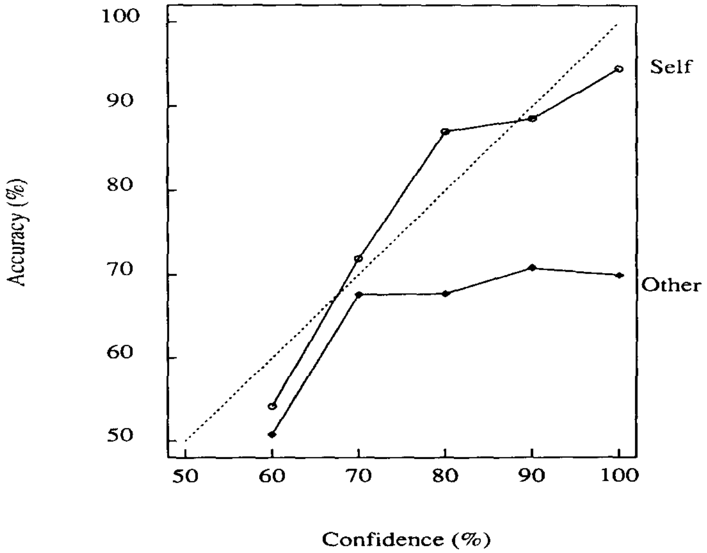

FIG. 5. Predicting self and other.

estimate the probability that they will choose one job over another. The average confidence in the predicted choice was a modest 66%, but only 1 of the 24 respondents chose the opinion to which he or she initially assigned a lower probability, yielding an overall accuracy rate of 96%. It is noteworthy that there are situations in which people exhibit overconfidence even in predicting their own behavior (Vallone, Griffin, Lin, & Ross, 1990). The key variable, therefore, is not the target of prediction (self versus other) but rather the relation between the strength and the weight of the available evidence.

The tendency to be confident about the prediction of the behavior of others, but not of one's own behavior, has intriguing implications for the analysis of decision making. Decision analysts commonly distinguish between decision variables that are controlled by the decision maker and state variables that are not under his or her control. The analysis proceeds by determining the values of decision variables (i.e., decide what you want) and assigning probabilities to state variables (e.g., the behavior of others). Some decision analysts have noted that their clients often wish to follow an opposite course: determine or predict (with certainty) the behavior of others and assign probabilities to their own choices. After all, the behavior of others should be predictable from their traits, needs, and interests, whereas our own behavior is highly flexible and contingent on changing circumstances (Jones & Nisbett, 1972).

The Effect of Difficulty

The preceding analysis suggests that people assess their confidence in one of two competing hypotheses on the basis of their balance of arguments for and against this hypothesis, with insufficient regard for the quality of the data. This mode of judgment gives rise to overconfidence when people form a strong impression on the basis of limited knowledge and to underconfidence when people form a moderate impression on the basis of extensive data.

The application of this analysis to general knowledge questions is complicated by the fact that strength and weight cannot be experimentally controlled as in Studies 1–3. However, in an analogy to a chance setup, let us suppose that the balance of arguments for a given knowledge problem can be represented by the proportion of red and white balls in a sample. The difficulty of the problem can be represented by the discriminability of the two hypotheses, that is, the difference between the probability of obtaining a red ball under each of the two competing hypotheses. Naturally, the greater the difference, the easier the task, that is, the higher the posterior probability of the more likely hypothesis on the basis of any given sample. Suppose confidence is given by the balance of arguments, that is, the proportion of red balls in the sample. What is the pattern of results predicted by this model?

Figure 6 displays the predicted results (for a sample of size 10) for three pairs of hypotheses that define three levels of task difficulty: an "easy" task where the probability of getting red balls under the competing hypotheses are respectively .50 and .40; a "difficult" task, where the probabilities are .50 and .45; and an "impossible" task, where the probability of drawing a red ball is .5 under both hypotheses. We have chosen non-symmetric hypotheses for our example to allow for an initial bias that is often observed in calibration data.

It is instructive to compare the predictions of this model to the results of Lichtenstein & Fischhoff (1977) who investigated the effect of task difficulty (see Fig. 7). Their "easy" items (accuracy = 85%) produced underconfidence through much of the confidence range, their "difficult" items (accuracy = 61%) produced overconfidence through most of the confidence range, and their "impossible" task (discriminating European from American handwriting, accuracy = 51%) showed dramatic overconfidence throughout the entire range.

A comparison of Figs. 6 and 7 reveals that our simple chance model reproduces the pattern of results observed by Lichtenstein & Fischhoff (1977): slight underconfidence for very easy items, consistent overconfidence for difficult items, and dramatic overconfidence for "impossible" items. This pattern follows from the assumption that judged confidence

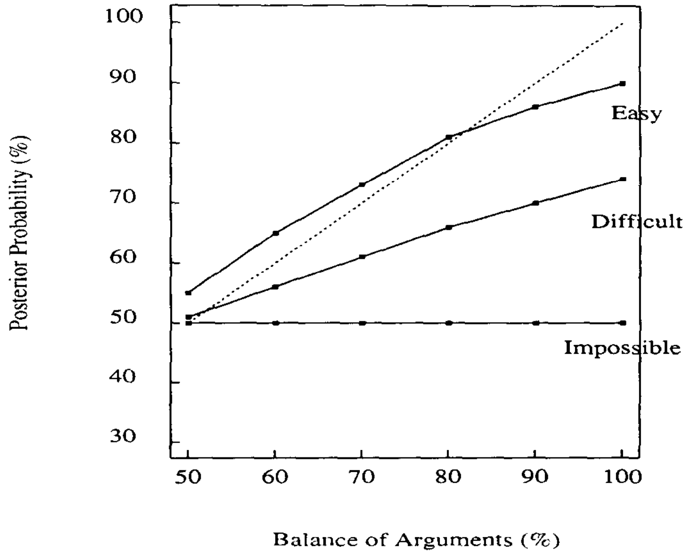

FIG. 6. Predicted calibration for item difficulty.

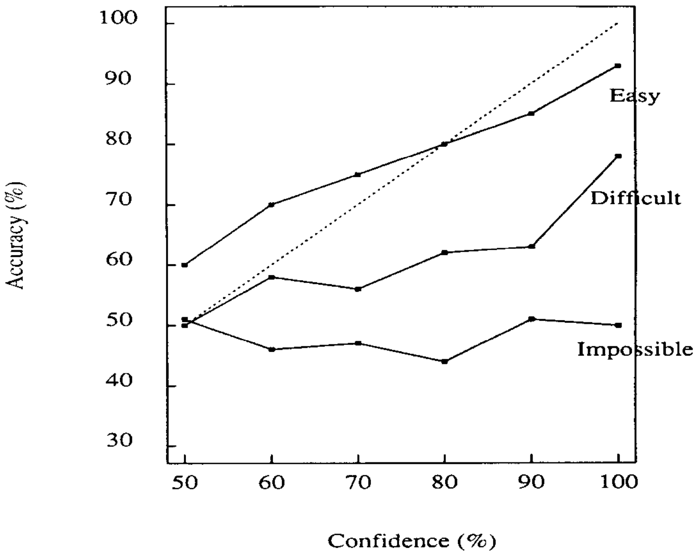

FIG. 7. Calibration plots for item difficulty.

is controlled by the balance of arguments for the competing hypotheses. The present account, therefore, can explain the observed relation between task difficulty and overconfidence (see Ferrell & McGoey, 1980).

The difficulty effect is one of the most consistent findings in the calibration literature (Lichtenstein & Fischhoff, 1977; Lichtenstein, Fischhoff, & Phillips, 1982; Yates, 1990). It is observed not only in general knowledge questions, but also in clinical diagnoses (Oskamp, 1962), predictions of future events (contrast Fischhoff & MacGregor, 1982, versus Wright & Wisudha, 1982), and letter identification (Keren, 1988). Moreover, the difficulty effect may contribute to other findings that have been interpreted in different ways. For example, Keren (1987) showed that world-class bridge players were well-calibrated, whereas amateur players were overconfident. Keren interpreted this finding as an optimism bias on the part of the amateur players. In addition, however, the professionals were significantly more accurate than the amateurs in predicting the outcome of bridge hands and the difference in difficulty could have contributed to the difference in overconfidence.

The difficulty effect can also explain the main finding of a study by Gigerenzer, Hoffrage, & Kleinbolting (1991). In this study, subjects in one group were presented with pairs of cities and asked to choose the city with the larger population and indicate their confidence in each answer. The items were randomly selected from a list of all large West German cities. Subjects in a second group were presented with general knowledge questions (e.g., Was the zipper invented before or after 1920?) and instructed to choose the correct answer and assess their confidence in that answer. Judgments about the population of cities were fairly well calibrated, but responses to the general knowledge questions exhibited overconfidence. However, the two tasks were not equally difficult: average accuracy was 72% for the city judgments and only 53% for the general knowledge questions. Hence, the presence of overconfidence in the latter but not in the former could be entirely due to the difficulty effect, documented by Lichtenstein & Fischhoff (1977). Indeed, when Gigerenzer et al. (1991) selected a set of city questions that were matched in difficulty to the general knowledge questions, the two domains yielded the same degree of overconfidence. The authors did not acknowledge the fact that their study confounded item generation (representative versus selective) with task difficulty (easy versus hard). Instead, they interpret their data as confirmation for their theory that overconfidence in individual judgments is a consequence of item selection and that it disappears when items are randomly sampled from some natural environment. This prediction is tested in the following study.

Study 5: The Illusion of Validity

In this experiment, subjects compared pairs of American states on several attributes reported in the *1990 World Almanac*. To ensure representative sampling, we randomly selected 30 pairs of American states from the set of all possible pairs of states. Subjects were presented with pairs of states (e.g., Alabama, Oregon) and asked to choose the state that was higher on a particular attribute and to assess the probability that their answer was correct. According to Gigerenzer et al. (1991), there should be no overconfidence in these judgments because the states were randomly selected from a natural reference class. In contrast, our account suggests that the degree of overconfidence depends on the relation between the strength and weight of the evidence. More specifically, overconfidence will be most pronounced when the weight of evidence is low and the strength of evidence is high. This is likely to arise in domains in which people can readily form a strong impression even though these impressions have low predictive validity. For example, an interviewer can form a strong impression of the quality of the mind of a prospective graduate student even though these impressions do not predict the candidate's performance (Dawes, 1979).

The use of natural stimuli precludes the direct manipulation of strength and weight. Instead, we used three attributes that vary in terms of the strength of impression that subjects are likely to form and the amount of knowledge they are likely to have. The three attributes were the number of people in each state (Population), the high-school graduation rate in each state (Education), and the difference in voting rates between the last two presidential elections in each state (Voting). We hypothesized that the three attributes would yield different patterns of confidence and accuracy. First, we expected people to be more knowledgeable about Population than about either Education or Voting. Second, we expected greater confidence in the prediction of Education than in the prediction of Voting because people's images or stereotypes of the various states are more closely tied to the former than the latter. For example, people are likely to view one state as more "educated" than another if it has more famous universities or if it is associated with more cultural events. Because the correlations between these cues and high-school graduation rates are very low, however, we expected greater overconfidence for Education than for Population or Voting. Thus, we expected high accuracy and high confidence for Population, low accuracy and low confidence for Voting, and low accuracy and higher confidence for Education.

To test these hypotheses, 298 subjects each evaluated half (15) of the pairs of states on one of the attributes. After subjects had indicated their confidence for each of the 15 questions, they were asked to estimate how many of the 15 questions they had answered correctly. They were reminded that by chance alone the expected number of correct answers was 7.5.

Table 4 presents mean judgments of confidence, accuracy, and estimated frequency of correct answers for each of the three attributes. Judgments of confidence exhibited significant overconfidence ($p < .01$) for all three attributes, contradicting the claim that "If the set of general-knowledge tasks is randomly sampled from a natural environment, we expect overconfidence to be zero" (Gigerenzer et al., 1991, p. 512). Evidently there is a great deal more to overconfidence than the biased selection of items.

The observed pattern of confidence and accuracy is consistent with our hypothesis, as can be seen in Fig. 8. This figure plots average accuracy against average confidence, across all subjects and items, for each of the

TABLE 4
Confidence and Accuracy for Study 6

	Population $N = 93$	Voting $N = 77$	Education $N = 118$
Confidence	74.7	59.7	65.6
Accuracy	68.2	51.2	49.8
Conf-Acc	6.5	8.5	15.8
Frequency	51.3	36.1	41.2

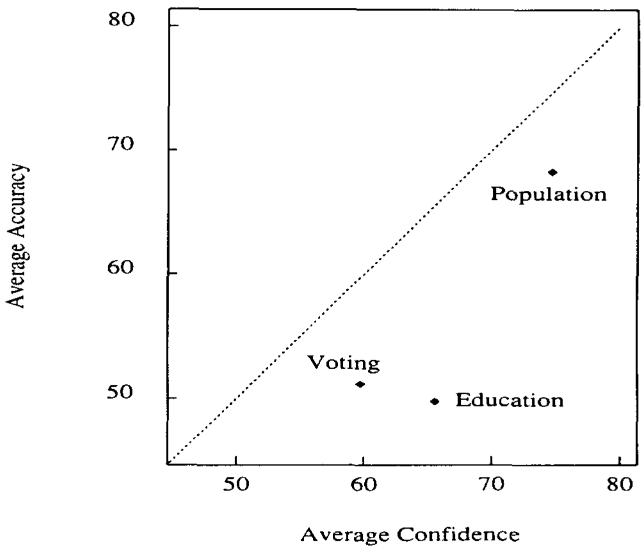

FIG. 8. Confidence and accuracy for three attributes.

three attributes. For Population, people exhibited considerable accuracy and moderate overconfidence. For Voting, accuracy was at chance level, but overconfidence was again moderate. For Education, too, accuracy was at chance level, but overconfidence was massive.

The present results indicate that overconfidence cannot be fully explained by the effect of difficulty. Population and Voting produced comparable levels of overconfidence (6.5 versus 8.5, $t < 1$, *ns*) despite a large difference in accuracy (68.2 versus 51.2, $p < .001$). On the other hand, there is much greater overconfidence in judgments about Education than about Voting (15.8 versus 8.5, $p < .01$) even though their level of accuracy was nearly identical (49.8 versus 51.2, $t < 1$, *ns*).

This analysis may shed light on the relation between overconfidence and expertise. When predictability is reasonably high, experts are generally better calibrated than lay people. Studies of race oddsmakers (Griffith, 1949; Hausch, Ziemba, & Rubinstein, 1981; McGlothlin, 1956) and expert bridge players (Keren, 1987) are consistent with this conclusion. When predictability is very low, however, experts may be more prone to overconfidence than novices. If the future state of a mental patient, the Russian economy, or the stock market cannot be predicted from present data, then experts who have rich models of the system in question are more likely to exhibit overconfidence than lay people who have a very limited understanding of these systems. Studies of clinical psychologists (e.g., Oskamp, 1965) and stock market analysts (e.g., Yates, 1990) are consistent with this hypothesis.

Frequency versus Confidence

We now turn to the relation between people's confidence in the validity of their individual answers and their estimates of the overall hit rate. A sportscaster, for example, can be asked to assess his confidence in the prediction of each game as well as the number of games he expects to predict correctly. According to the present account, these judgments are not expected to coincide because they are based on different evidence. A judgment of confidence in a particular case, we propose, depends primarily on the balance of arguments for and against a specific hypothesis, e.g., the relative strength of two opposing teams. Estimated frequency of correct prediction, on the other hand, is likely to be based on a general evaluation of the difficulty of the task, the knowledge of the judge, or past experience with similar problems. Thus, the overconfidence observed in average judgments of confidence need not apply to global judgments of expected accuracy. Indeed, Table 4 shows that estimated frequencies were substantially below the actual frequencies of correct prediction. In fact, the latter estimates were below chance for two of the three attributes.[3] Similar results have been observed by other investigators (e.g., Gigerenzer et al., 1991; May, 1986; Sniezek & Switzer, 1989). Evidently, people can maintain a high degree of confidence in the validity of specific answers even when they know that their overall hit rate is not very high.[4] This phenomenon has been called the "illusion of validity" (Kahneman & Tversky, 1973): people often make confident predictions about individual cases on the basis of fallible data (e.g., personal interviews or projective tests) even when they know that these data have low predictive validity (Dawes, Faust, & Meehl, 1989).

The discrepancy between estimates of frequency and judgments of confidence is an interesting finding but it does not undermine the significance of overconfidence in individual items. The latter phenomenon is important because people's decisions are commonly based on their confidence in their assessment of individual events, not on their estimates of their overall hit rate. For example, an extensive survey of new business owners (Cooper, Woo, & Dunkelberg, 1988) revealed that entrepreneurs were, on average, highly optimistic (i.e., overconfident) about the success of their specific new ventures even when they were reasonably realistic about the general rule of failure for ventures of that kind. We suggest that decisions to undertake new ventures are based primarily on beliefs about individual events, rather than about overall base rates. The tendency to prefer an

[3] One possible explanation for this puzzling observation is that subjects reported the number of items they knew with certainty, without correction for guessing.

[4] This is the statistical version of the paradoxical statement "I believe in all of my beliefs, but I believe that some of my beliefs are false."

individual or "inside" view rather than a statistical or "outside" view represents one of the major departures of intuitive judgment from normative theory (Kahneman & Lovallo, 1991; Kahneman & Tversky, 1982).

Finally, note that people's performance on the frequency task leaves much to be desired. The degree of underestimation in judgments of frequency was comparable, on average, to the degree of overconfidence in individual judgments of probability (see Table 4). Furthermore, the correlation across subjects between estimated and actual frequency was negligible for all three attributes (+.10 for Population, −.10 for Voting, and +.15 for Education). These observations do not support the view that people estimate their hit rate correctly, and that the confidence–frequency discrepancy is merely a manifestation of their inability to evaluate the probability of unique events. Research on overconfidence has been criticized by some authors on the grounds that it applies a frequentistic criterion (the rate of correct prediction) to a nonfrequentistic or subjective concept of probability. This objection, however, overlooks the fact that a Bayesian expects to be calibrated (Dawid, 1982), hence the theory of subjective probability permits the comparison of confidence and accuracy.

CONCLUDING REMARKS

The preceding study demonstrated that the overconfidence observed in calibration experiments is not an artifact of item selection or a byproduct of test difficulty. Furthermore, overconfidence is not limited to the prediction of discrete events; it has consistently been observed in the assessment of uncertain quantities (Alpert & Raiffa, 1982).

The significance of overconfidence to the conduct of human affairs can hardly be overstated. Although overconfidence is not universal, it is prevalent, often massive, and difficult to eliminate (Fischhoff, 1982). This phenomenon is significant not only because it demonstrates the discrepancy between intuitive judgments and the laws of chance, but primarily because confidence controls action (Heath & Tversky, 1991). It has been argued (see e.g., Taylor & Brown, 1988) that overconfidence—like optimism—is adaptive because it makes people feel good and moves them to do things that they would not have done otherwise. These benefits, however, may be purchased at a high price. Overconfidence in the diagnosis of a patient, the outcome of a trial, or the projected interest rate could lead to inappropriate medical treatment, bad legal advice, and regrettable financial investments. It can be argued that people's willingness to engage in military, legal, and other costly battles would be reduced if they had a more realistic assessment of their chances of success. We doubt that the benefits of overconfidence outweigh its costs.

REFERENCES

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). Cambridge: Cambridge University Press.

Bar-Hillel, M. (1983). The base rate fallacy controversy. In R. W. Scholz (Ed.), *Decision making under uncertainty* (pp. 39–61). Amsterdam: North-Holland.

Camerer, C. (1990). Do markets correct biases in probability judgment? Evidence from market experiments. In L. Green & J. H. Kagel (Eds.), *Advances in behavioral economics,* Vol. 2 (pp. 126–172).

Cooper, A. C., Woo, Carolyn, Y., & Dunkelberg, W. C. (1988). Entrepreneurs' perceived chances for success. *Journal of Business Venturing,* **3,** 97–108.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist,* **34,** 571–582.

Dawes, R. M. (1988). *Rational choice in an uncertain world.* New York: Harcourt Brace Jovanovich.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science,* **243,** 1668–1674.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association,* **77,** 605–613.

Dun & Bradstreet. (1967). *Patterns of success in managing a business.* New York: Dun and Bradstreet.

Dunning, D., Griffin, D. W., Milojkovic, J., & Ross. L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology,* **58,** 568–581.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York: Wiley.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance,* **26,** 32–53.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). New York: Cambridge.

Fischhoff, B., & Bar-Hillel, M. (1984). Focusing techniques: A shortcut to improving probability judgments? *Organizational Behavior and Human Performance,* **34,** 175–194.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review,* **90,** 239–260.

Fischhoff, B., & MacGregor D. (1982). Subjective confidence in forecasts. *Journal of Forecasting,* **1,** 155–172.

Forer, B. (1949). The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology,* **44,** 118–123.

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., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review,* **98,** 506–528.

Grether, D. M. (1980). Bayes' rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics,* **95,** 537–557.

Grether, D. M. (1990). *Testing Bayes' rule and the representativeness heuristic: Some experimental evidence* (Social Science Working Paper 724). Pasadena, CA: Division of the Humanities and Social Sciences, California Institute of Technology.

Griffin, D. W. (1991). *On the use and neglect of base rates.* Unpublished manuscript, Department of Psychology, University of Waterloo.

Griffith, R. M. (1949). Odds adjustments by American horse-race bettors. *American Journal of Psychology,* **62,** 290–294.

Hausch, D. B., Ziemba, W. T., & Rubinstein, M. (1981). Efficiency of the market for racetrack betting. *Management Science,* **27**, 1435–1452.

Heath, F., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty,* **4**, 5–28.

Jones, E. E., & Nisbett, R. E. (1972). *The actor and the observer: Divergent perceptions of the causes of behavior.* Morristown, NJ: General Learning Press.

Kahneman, D., & Lovallo, D. (1991). Bold forecasting and timid decisions: A cognitive perspective on risk taking. In R. Rumelt, P. Schendel, & D. Teece (Eds.), *Fundamental issues in strategy.* Cambridge: Harvard University Press, forthcoming.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology,* **3**, 430–454.

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

Kahneman, D., & Tversky, A. (1982). Intuitive prediction: Biases and corrective procedures. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 414–421). Cambridge: 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.

Kidd, J. B. (1970). The utilization of subjective probabilities in production planning. *Acta Psychologica,* **34**, 338–347.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory,* **6**, 107–118.

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., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge: Cambridge University Press.

Lusted, L. B. (1977). *A study of the efficacy of diagnostic radiologic procedures: Final report on diagnostic efficacy.* Chicago: Efficacy Study Committee of the American College of Radiology.

McGlothlin, W. H. (1956). Stability of choices among uncertain alternatives. *American Journal of Psychology,* **69**, 604–615.

May, R. S. (1986). Inferences, subjective probability and frequency of correct answers: A cognitive approach to the overconfidence phenomenon. In B. Brehmer, H. Jungermann, P. Lourens & G. Sevo'n (Eds.), *New directions in research on decision making* (pp. 175–189). Amsterdam: North-Holland.

Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest,* **2**, 2–9.

Neale, M. A., & Bazerman, M. H. (1990). *Cognition and rationality in negotiation.* New York: The Free Press, forthcoming.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of human judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs,* **76** (28, Whole, No. 547).

Oskamp, S. (1965). Overconfidence in case-study judgments. *The Journal of Consulting Psychology,* **29,** 261-265.

Peterson, C. R., & Miller, A. J. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology,* **70,** 117-121.

Peterson, C. R., Schneider, R. J., & Miller, A. J. (1965). Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology,* **69,** 522–527.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology,* **72,** 346–354.

Quattrone, G. A. (1982). Overattribution and unit formation: When behavior engulfs the person. *Journal of Personality and Social Psychology,* **42,** 593–607.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance,* **6,** 649–744.

Sniezek, J. A., & Switzer, F. S. (1989). *The over-underconfidence paradox: High Pi's but poor unlucky me.* Paper presented at the Judgment and Decision Making Society annual meeting in Atlanta, Georgia.

Staël von Holstein, C.-A. S. (1972). Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance,* **8,** 139–158.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin,* **103,** 193–210.

Tversky, A., & Kahneman, D. (1971). The belief in the law of small numbers. *Psychological Bulletin,* **76,** 105–110.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science,* **185,** 1124–1131.

Vallone, R. P., Griffin, D. W., Lin, S., & Ross, L. (1990). The overconfident prediction of future actions and outcomes by self and others. *Journal of Personality and Social Psychology,* **58,** 582-592.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research.* New York: Cambridge University Press.

Wagenaar, W. A., & Keren, G. (1986). Does the expert know? The reliability of predictions and confidence ratings of experts. In E. Hollnagel, G. Maneini, & D. Woods (Eds.), *Intelligent decision support in process environments* (pp. 87–107.) Berlin: Springer.

Wright, G., & Wisudha, A. (1982). Distribution of probability assessments for almanac and future event questions. *Scandinavian Journal of Psychology,* **23,** 219–224.

Yates, J. F. (1990). *Judgment and Decision Making.* Englewood Cliffs, NJ: Prentice–Hall.

(Accepted October 30, 1991)