Acta Psychologica 77 (1991) 217–273
North-Holland

Calibration and probability judgments: Conceptual and methodological issues

Gideon Keren *
University of Utrecht, Utrecht, The Netherlands

Accepted April 1991

In a world characterized by uncertainty, the study of how people assess probabilities carries both theoretical and practical implications. Much of the research efforts in this area, especially in psychology, has focused on calibration studies (Lichtenstein, Fischhoff and Phillips 1982). The present article offers an extensive review of conceptual and methodological issues involved in the study of calibration and probability assessments. It is claimed that most calibration studies have focused on technical formal issues and are in this respect a-theoretical. The reason for this state of affairs is the adoption of a *strict* perspective which assumes that uncertainty is a reflection of the external world, and relies heavily on normative and formal considerations. Several unresolved problems within this strict outlook are pointed out. The present paper assumes that calibration (and assessments of subjective probabilities in general) is not a characteristic of the event(s), but rather of the assessor (Lad 1984), and advocates a more *loose* perspective, which is broader and more descriptive in nature. Possible discrepancies between a strict and a more loose perspective, as well as reconciliation attempts, are presented.

Dealing with uncertainty is an important component of human behavior. Coping with uncertain knowledge, predicting likely outcomes of decision choices, assessing personal risks, judging the motives and actions of others, and recalling past experience are everyday events that require explicit or implicit probability judgments. At the societal level, such judgments are of importance in policy decisions, risk assessments, crime investigation and trials, medical diagnosis, and many other areas.

There exists a large body of research on probability judgments that is interdisciplinary in nature, summarized in several review articles (e.g., Hogarth, 1975; Hampton et al. 1973; Spetzer and Stael von Holstein 1975; Wallsten and Budescu 1983). A fundamental question

* Requests for reprints should be sent to G. Keren, Dept. of Psychology, Free University of Amsterdam, De Boelelaan 1115, 1081 HV Amsterdam, The Netherlands.

0001-6918/91/$03.50

concerning probability judgments is how should they be assessed and evaluated? Underlying this question are two related facets, a normative and a descriptive.

It is important to realize that normative and descriptive aspects of probability judgments are deeply intertwined. Probability judgments are means for conveying *information* about uncertainty. From a normative perspective the question is how could this information be conveyed in a most accurate, coherent, and efficient way? A related subsequent question then arises concerning the proper evaluation standards that should be used in order to assess whether these criteria have been satisfied? Winkler and Murphy (1968) suggested two such standards: *normative goodness* which is an assessment of the degree to which probability judgments truly reflect the assessors' beliefs and conform to the axioms of probability theory; and *substantive goodness* which is supposed to reflect assessors' knowledge of the particular domain in which the probabilistic judgments are being made. A third criterion, by far the most commonly used, is *calibration* which may be conceived as an accuracy measure. It evaluates the fit between probability judgments and the corresponding events to which they refer.

From a descriptive viewpoint, the question is whether actually the appropriate information has been accurately and reliably transmitted? Again, the problem of what are the proper evaluation criteria that should be used, is being raised. [1] The common way to answer the above question is to test whether actual probability judgments (and their interpretation) conform and satisfy the normative criteria. If they do not, the question is whether they deviate in a systematic and predictable manner, and why?

Any discrepancy between the actual probability judgments and the normative yardstick can be interpreted in several different ways, which could be classified in three categories: (i) The discrepancy could be due to the fact that intuitive judgments are simply incompatible with the ideal normative theory. The account for such discrepancies would then be mainly in terms of cognitive processes underlying probability judgments. Indeed, one major goal of descriptive research is to uncover the nature of such processes. Alternatively, the source of such discrepancies could be due to either (ii) normative evaluation criteria that are

[1] It is this problem of identifying the appropriate evaluation criteria where normative and descriptive considerations are most interrelated and cannot be completely separated.

ambiguous or lack a sound justification, or (iii) inaccurate interpretation of the actual probability judgments, due to inappropriate elicitation methods, misleading procedures, or other methodological inadequacies. Explanations based on these three types are often confounded and difficult to separate. The present paper provides a systematic analysis of conceptual issues associated with (ii), and critically reviews procedural and methodological aspects that are related to (iii). The purpose is to facilitate the separation between the three categories, and at the same time provide the grounds for developing an adequate behavioral theory of probabilistic judgments (which alludes to (i)) that would be uncontaminated by factors associated with categories (ii) and (iii).

The focus of the present paper is on calibration. Strictly speaking, calibration is a criterion for evaluation of probability judgments (e.g., Baron, 1988), and as such is part of the normative arsenal. Much of the research effort in this area, however, has centered on the psychological and descriptive aspects (for an extensive review, see Lichtenstein et al. 1982). With few exceptions (e.g., Koriat et al. 1980; Pitz 1974), however, most of this research has been technical in nature, lacks an adequate theoretical framework (Lichtenstein et al. 1982), and empirical results are usually accounted for by some post-hoc explanations.

A major reason for this lack of theorizing stems from confusions resulting from the interrelationship between the normative and the descriptive facets, and from unresolved conceptual and methodological difficulties (categories (ii) and (iii) above) that arise when evaluating probability assessments. The purpose of this article is to provide a systematic review of these conceptual and methodological issues in the context of calibration studies.

The fundamental assumption underlying the present paper is that calibration (and assessments of subjective probabilities in general) is not a characteristic of the event(s), but rather of the assessor (Lad 1984), and as such is a proper theme for psychological investigation. Consequently, our review of conceptual and methodological issues is strongly linked with the psychological and descriptive facets underlying probability judgments.

In the first part of the paper, the main experimental paradigm employed in calibration studies is described, and a brief summary of the major empirical findings is presented. In the second section we discuss conceptual issues regarding the interpretation of probability

and calibration. We point out several normative problems that have not been completely resolved, and claim that these complicate the interpretation of descriptive data. The third section is devoted to methodological matters regarding procedures of eliciting subjective probabilities. Since probability estimates are not invariant across different elicitation methods, care should be taken in the interpretation of empirical results obtained by different methods. In the fourth section we present assessors' characteristics such as goals, motivation, experience, and loss functions, that are relevant to probability judgment and suggest that the study of such aspects is essential for the development of an adequate descriptive theory of probability judgment. The last section deals with formal measures of calibration. Although this issue is normative in nature, it has shaped in many respects the direction and interpretation of empirical studies.

It is apparent from our review that several conceptual and methodological issues remain unresolved, thus often leading to ambiguous and equivocal interpretations of empirical results. In the final concluding section it is claimed that a major source of the ostensible difficulties is due to an uncompromising reliance on a strict normative framework that was embraced by most researchers. A looser framework is advocated that (i) enables to take into account different interpretations of probability, depending on the nature of the particular task and stimulus material, (ii) in which probability judgments are assessed not just by the final outcome but also by the process by which these were derived, and(iii) centers on the cognitive mechanisms underlying probability judgments.

Calibration studies: The paradigm and major empirical results

The experimental paradigm

Calibration studies are concerned with the appropriateness of assessors' subjective probability estimates, or confidence in their judgments and predictions, and can be categorized in two groups: one that elicits judgments about discrete propositions, and one that attempts to identify probability density functions assessed over continuous variables (e.g., uncertain numerical values).

The customary definition for discrete probability statements is that

judgments are well calibrated 'if on the long run, for all propositions assigned a given probability, the proportion that is true is equal to the probability assigned' (Lichtenstein et al. 1982). Discrete probability statements can be classified according to the number of possible alternatives the assessor is exposed to, and the corresponding range of the probability scale: in the *one alternative case*. the assessor is required to make a probability judgment with regard to a single event or statement (provided either by the assessor or by someone else). The appropriate probability response in this case ranges between 0 and 1.0. In the *two alternatives case* the assessor has to choose between two alternatives, and then provide a probability judgment for the chosen alternative in the range of 0.5 to 1.0. Finally, in the *multiple alternatives* case, the assessor is asked to select the most likely response among K options, followed by a probability estimate in the range $1/K$ to 1.0. Since the majority of the studies employed the two alternatives case, and for the purpose of parsimony, most of the discussion in the present paper will focus on this case. However, unless otherwise stated, the conclusions are readily generalizable to other conditions.

A common way of analyzing probability judgments and confidence ratings is via the use of a calibration curve in which the hit rate (percent correct) is plotted on the ordinate for each confidence response plotted on the abscissa. It is customary to group all responses in the range 0.50–0.59, 0.60–0.69,...,0.90–0.99, and 1.0. The mean percentage correct for each response group is then plotted against the corresponding mean probability assessment for that category. The 45 deg line represents perfect calibration. Any point below this line is interpreted as reflecting overconfidence, and any point above it represents underconfidence. Under- or overconfidence can be further assessed by the weighted mean (over response groups) of the differences between the mean of the probability responses and the corresponding proportion correct for each category (Lichtenstein and Fischhoff 1977) or:

$$1/N * \sum n_t(r_t - c_t), \tag{1}$$

where N stands for the total number of responses, t the number of response categories, n_t the number of times the response r_t has been used, and c_t the corresponding proportion of correct items. Any positive result of the above measure indicates overconfidence (the

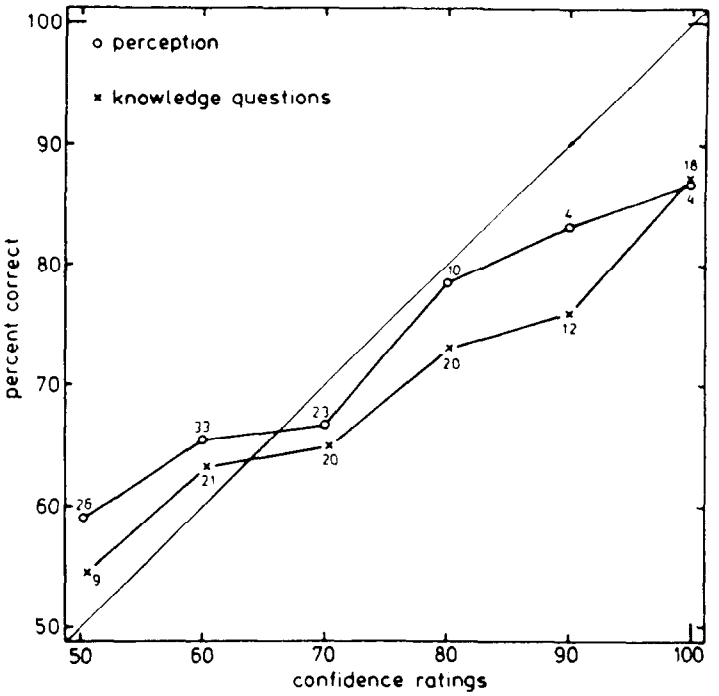

Fig. 1. Calibration curves for a perceptual task and for general knowledge questions based on Keren (1988a).

larger the number, the larger the overconfidence), and a negative score implies underconfidence.

As an example, Fig. 1 depicts two calibration curves from a recent study by Keren (1988a). The lower curve represents the results of a general-knowledge task and shows a strong tendency of overconfidence. The upper curve is based on a perceptual task showing slight underconfidence (as computed by eq. (1)). Note that the numbers adjacent to each point (on either curve) represent relative frequencies of observations expressed in terms of percentages. This is important since the deviations of a curve from the 45 deg line (indicating the extent of over- underconfidence) should be weighted by the relative number of observations for each point on the curve. For example, inspecting the curve of the perceptual task in Fig. 1, overconfidence is apparent for confidence ratings of 100% but these account only for 4% of all the observations. In contrast, the points corresponding to confidence ratings of 50% and 60% show underconfidence, and account for

26% and 33% of the observations, respectively. Unfortunately, most articles that present calibration curves have failed to follow this practice.

Calibration regarding uncertain continuous quantities can be assessed by estimating the probability distribution (over the continuum) with the use of different fractiles. Roughly, calibration is intended to measure the extent to which a set of probability density functions 'corresponds to reality'. Over- or underconfidence are usually measured by the *interquartile* index, and the *surprise index*. The former is the percentage of items for which the true value (actual outcome) falls inside the interquartile range (i.e., between the 0.25 and the 0.75 fractiles), and perfect calibration is indicated by an index of 50. Any value lower than 50 would imply overconfidence, and values above 50 are interpreted as underconfidence. The surprise index represents the percentage of true values falling outside the most extreme fractiles assessed. For instance, a surprise index of 2% refers to the extreme values that fall below 0.01 and above the 0.99 fractiles. Ideal calibration would lead to a surprise index of 2, and any value above it would represent overconfidence. Frequently, the relative frequency of true values falling below the assessed medians is also computed (Seaver et al. 1978).

Major empirical findings [2]

Two major consistent findings emerge from the calibration literature. One is a pervasive tendency of overconfidence, in particular (but not restricted to) in tasks that use general-knowledge questions. Notable exceptions are some studies with experts that are further discussed below. The second main finding is that the degree of overconfidence (when it exists) depends on the difficulty of the task as measured by the percent of correct responses or predictions. Usually, the more difficult the task the larger the degree of overconfidence as indicated by the traditional measures. Whether this finding is psychologically meaningful and not a result of some unnoticed artifacts, is further analyzed in subsequent sections.

One question, of particular interest, is whether experts (in any given

[2] This section is intentionally brief and provides only a short summary of those issues that are essential for the following discussion. A comprehensive review of the methods used in calibration studies and a broader review of empirical results can be found in Lichtenstein et al. (1982). Additional measures of calibration not discussed in this part, are presented in a later section.

field) are better calibrated compared with lay people. Put differently, is good calibration a necessary attribute of expertise? The answer to this question is not unequivocal: some studies reported high overconfidence, especially for different types of diagnosis in the medical field (e.g. Christensen-Szalanski and Busheyhead 1981; Fryback and Erdman 1979; Lusted 1977). Chan (1982) reviewed various studies with experts and suggested that in several cases experts' probability assessments were not better than those of lay persons. In contrast, other studies in different fields resulted in good calibration, showing little, if any, overconfidence (occasionally even some underconfidence). This was true for weather forecasters (Murphy and Winkler 1977), accountants (Tomassini et al. 1982), professional bridge players (Keren 1987), and students predicting their course grade (Sieber 1974). The last finding has been recently challenged by Manger and Teigen (1988). They report a high level of overconfidence of students predicting their grades, under different time horizons (i.e., eight and two months before their final exam).

These mixed results raise several questions: What is the source of these large differences between experts in making appropriate probability judgments? Is it due to what Murphy and Winkler termed normative goodness? And if so, is there a general skill involved in making well-calibrated probability judgments? Or perhaps, as many claim, substantive knowledge is the decisive factor. But if the latter is the case, why are experts in some domains better calibrated than experts in other domains? The question may be raised whether the nature of stimulus material and characteristics of the prediction task may account for the different performance of different expert groups. These are main theoretical questions for which the current literature provides only tentative and partial answers.

Interpreting calibration studies: Conceptual issues

Calibration and the meaning of subjective probabilities

The most fundamental problem underlying the evaluation of probabilistic assessments, in particular the interpretation of calibration studies, stems from the multiple meanings that can be attached to a probability statement. Avoiding a lengthy discussion on the different meanings and possible interpretations of a probability statement (e.g.,

Kyburg 1983; Kyburg and Smokler 1964; Lindley et al. 1979; Salmon 1967), it is important to point out an inherent problem in the literature on calibration studies: on the one hand, these studies assume, implicitly or explicitly, that probability statements are purely subjective and hence represent a mere internal state (Lichtenstein et al. 1982). It is for this reason that such statements cannot be judged as 'right' or 'wrong'. On the other hand, evaluation and analysis of these studies is virtually always based on a frequentistic interpretation of probability. Thus, probabilistic judgments (which are supposedly subjective) are said to be well calibrated when the relative frequencies of events match the corresponding judged probabilities. How can this inconsistency be resolved?

The incompatibility between subjective probabilities and the assessment of their accuracy has been recognized almost since the concept of probability first emerged. Hacking (1975), in his excellent review of the history of probability, noticed that even as early as the seventeenth century the concept of probability was already associated with two distinct facets. One, which he termed *aleatory*, concerns itself with stochastic laws of chance processes and stable frequencies (on the long run), and the other, which he termed *epistemological*, refers to the degree of belief and is in essence devoid of statistical considerations. This *duality* problem, as Hacking refers to it, is evident through the entire history of probability since it was conceived, and in fact has not been resolved until today.

Lindley et al. (1979) made a reconciliation attempt by proposing to distinguish between different types of criteria for the evaluation of probability assessments. Calibration, which they termed the *semantic criterion*, pertains to the meaning of the probability scale. Lindley et al., were aware, however, that a prerequisite for applying this criterion in a meaningful way is a sizeable number of observations, and reliance on a frequentistic interpretation that is incompatible with the subjective view. They thus state that 'there is no way of validating, for example, a meteorologist's single judgment that the probability of rain is 2/3. If the meteorologist is using the scale properly, however, we would expect that rain would occur on about two-thirds of the days which he assigns a probability of 2/3' (p. 147). [3] Later on, Lindely et al. actually suggest a more relaxed view for assessing subjective probabilities though the

[3] Fischhoff and MacGregor (1982) also noted that unique predictions cannot be validated except for the special case in which an individual is 100% confident (or certain) and wrong.

only justification for it is on practical and pragmatic grounds supported by face validity.

Harrison (1977) warns 'that the widespread belief that assessed probability and observed frequency should agree seems to be vaguely based on the law of large numbers, but this reflects a basic misunderstanding' (p. 324). The source of the misunderstanding, according to Harrison, is reliance on the law of large numbers that encapsulates the independence concept. However, Harrison claims, in the subjective interpretation there is no concept of independence that is external to the assessor, and thus there is no justification for using a frequentistic criterion.

A lucid formulation of the subjectivists school has been offered by Lad (1984). He claims that 'since probabilities (previsions) are not predictions, there is no point in comparing the previsions with outcomes as if it made sense to ask whether they had been correct or not. Posterior probabilities of subsequent events are not corrections of previous probabilities. They are simply a new evaluation, cohering with the previous one, corresponding to the new state of information' (p. 215).

One of the most well-known attempts to (formally) incorporate observed frequencies into the subjective perspective (though for a restrictive set of events) has been proposed by De Finetti who introduced the concept of *exchangeability* (e.g., De Finetti 1970). A sequence of events is said to be exchangeable if 'the probability of any particular distribution of a given property in a finite set of events of the sequence depends only on the *number* of events in that distribution that have that property' (Kyburg 1968: 68). In other words, events of a set are exchangeable if each single event, pair of events, triples (and so forth) in the set have the same probability. An obvious and intuitive example of exchangeable events are drawings from an urn with an unknown composition of white and black balls, and more generally any case of 'repeated trials with a constant but unknown probability of success' (De Finetti 1970).

It is important to emphasize that exchangeability is an entirely subjective notion. Yet, as shown by De Finetti, events that are exchangeable lend themselves to conform to the observed frequencies of the corresponding events. Consequently, according to De Finetti, there is a justification to evaluate probabilistic estimates by a frequentistic criterion provided that the underlying events are exchangeable. Note,

however, that exchangeability does not ensure good calibration (Kadane and Lichtenstein 1982).

In summary, despite the continuous reconciliation efforts, few that were briefly discussed above, none in fact offer a genuine satisfactory solution. From a strict formal viewpoint, it is apparent that the duality problem (Hacking 1975) is deeply rooted in the concept of probability, and the internal incompatibility between the two facets has never been completely resolved. The pragmatic solution, adopted in virtually all calibration studies, has been to overlook the inconsistency and appraise subjective probabilities by frequentistic means.

Related items and similarity of events in the set

The notion of exchangeable events offers a normative criterion for judging when a set of probability judgments can be evaluated by a frequentistic criterion. Exchangeability, however, is entirely based on formal distributional properties that are external to the assessor: the number of events in the set and the equal probability of their occurrence. The approach adopted in this article is that calibration of subjective probabilities is mainly a characteristic of the assessor (rather than the events), and should therefore be reflected in any attempt to reconcile the subjective view with a frequentistic interpretation.

Both Keren (1987), and Ronis and Yates (1987) pointed out that even if a sufficiently large number of repeated predictions (i.e., probability judgments) can be obtained, there may still be a fundamental difference between the nature of different tasks. In particular, although every event literally occurs only once, we can nevertheless identify two different conditions: in one, the set of predicted events are what Ronis and Yates (1987) called *essentially similar*, and referred to by Keren (1987) as *related* items. Related items share common characteristics and consequently the information relevant to the probability assessment of any event in such a set is drawn from the same data base. [4]

[4] The notion of related or essentially similar items is not entirely new. Von Mises (1957) proposed that a probability statement is meaningful only if one specifies the relevant *collective*. The collective is defined as 'a sequence of uniform events or processes which differ by certain observable attributes' (p. 12). While the notion of a collective is close to a set of related items, it is important to emphasis that von Mises' concept was developed within a highly restrictive frequentistic theory. In contrast, the emphasis in introducing related items is on the subjective (cognitive) mechanisms that determine the appropriate set. As such, the latter serves as an attempt to bridge between the frequentistic (objective) and subjective interpretations of probability.

Weather forecasts or medical diagnoses of a certain disease constitute essentially similar or related items, and thus outcome knowledge of previous events can be incorporated in a data base that may be useful for subsequent predictions. In contrast, events that do not share primary and vital (stimulus) characteristics are said to be *essentially unique* or *unrelated*. For instance, general-knowledge questions like 'which country has a larger population: Peru or Norway?' and 'who died first: Mozart or Schubert?' constitute unrelated items, since subjects' knowledge of one item is completely independent of their knowledge of the other item. Each item in an unrelated set (or derived from a different data base) constitutes its own knowledge domain, and no information or knowledge can be inferred or obtained from other items. In addition, as Ronis and Yates (1988) pointed out, 'The alternatives in general-knowledge questions also constitute essentially unique "events" ' (p. 200).

Primarily, what matters is whether the diverse events belong to the same *set* or sample space (as perceived by the assessor), so that they can be viewed as 'replications' of each other. Replications in this context are not limited to 'statistical' replications, but to events or items which, because of their perceived similarity, are supposedly handled by the same cognitive processes. In that respect, there is some similarity between psychophysics and calibration studies.

The goal of psychophysics is to discover relationships between stimuli in the external world and the corresponding internal states (usually referred to as sensations) produced by these stimuli. The internal state (assumed to be reflected in the person's response) is always assessed against a well-defined criterion, namely the outer physical world. Calibration studies also attempt to assess an internal state (i.e., degree of uncertainty), but the criterion against which they are matched is not directly provided by the external world.

Two underlying assumptions of psychophysics are that (i) the physical stimuli presented on successive trials are (for all practical purposes) the *same*, and thus are justified to serve as *replications*, and that consequently (ii) the perceptual mechanisms employed for encoding and processing of these stimuli are also the same on successive trials. In calibration studies, the justification for assumption (i) is based more on subjective judgments of similarity and relatedness, because a well-defined external criterion is not available. Consequently, also the second assumption may be weaker with regard to calibration compared with

psychophysics studies. Using our terminology, stimuli in psychophysics satisfy the most extreme case of essentially similar items. Keren (1988a) has demonstrated that subjects are well calibrated when psychophysical tasks are involved, whereas the same subjects were poorly calibrated with general knowledge questions (the difficulty of the two tasks, as measured by percent correct responses, was the same in both tasks).

Related sets may be 'naturally' formed or alternatively, under laboratory conditions, may be created by an act of the experimenter. In the latter case, it becomes important to know whether this has been made clear to subjects, and the degree to which subjects accepted the definition imposed on them. Even if accepted, a set of events may be so remote from subjects' ordinary way of organizing their experiences (their mental model of the world) that they may not be able to readily calculate their chances of being correct from past successes and failures. It is therefore important to distinguish between potential similarity (based on identifying certain common attributes) that may seem to satisfy the requirements for relatedness, and the actual perceived similarity as experienced by the assessor (or the subject in a controlled experiment). [5]

Kahneman and Tversky (1982) offered a distinction based on the nature of the data that a judge may consider in the process of estimating probabilities, which is closely connected to the notion of relatedness. Uncertainty can be assessed in a *distributional* mode in which the case in question can be considered as an instance of a class of similar events, and for which the relative frequencies of outcomes are either known or can be estimated. On the other hand, instances that cannot be classified as similar events, are processed in a *singular* mode in which probabilities are said to be estimated by propensities of the particular case at hand. Whereas related items lend themselves naturally to the distributional mode of processing, unrelated items can only be processed in a singular mode.

The distinction between related (similar) and unrelated (unique) events is important for two reasons: first, it provides a valid justifica-

[5] This distinction is based on Wallach's (1956) suggestion, that common attributes displayed by objects may give rise to similarity judgments leading to what he termed *potential* similarity. As he noticed, however, 'The attributive basis for a judgment of similarity may be present and yet the judgment not be made; and, on the other hand, an attributive basis for such a judgment may be lacking and yet the items may be judged similar. There would seem to be a difference, then, between potential similarity and psychological similarity' (p. 103).

tion for using a frequentistic criterion for assessing calibration, whenever related or essentially similar events are concerned. It would make little sense, however, to apply it to unrelated events: that would be analogous to adding apples and pears and averaging across them.

A second reason, mainly from a descriptive view point, is related to the alleged cognitive processes that may be employed for related and unrelated events. Consider first the condition of essentially unique events, such as general-knowledge questions. The statement 'I am 80% certain that Ankara is the capital of Turkey' can be interpreted as equivalent to the following two statements: (1) Ankara is the capital of Turkey, and (2) I am 80% certain that this statement is correct. It is often assumed that these two separate statements reflect a serial process: first a decision (or choice between alternatives) is made that Ankara is the capital of Turkey, followed by a probabilistic assessment regarding the truth of this assessment.

Next, consider the case of essentially similar events such as weather forecasting. The statement 'there is 80% chance of rain tomorrow' can also be interpreted, as in the unique case, to imply two separable statements (again assumed to be processed in a serial manner) namely: (1) It will rain tomorrow, and (2) I am 80% certain that this statement is correct. An alternative hypothesis in the case of related events, and one that is even more plausible, is based on a frequentistic interpretation. According to this account, the weather forecaster conveys to us that in the past, under identical or very similar weather conditions, there was rain in approximately 80% of the cases; hence, the probability that it will rain tomorrow is 0.8. This interpretation assumes that the to-be-predicted event is first classified with a set of similar previous events, and that the probability estimate is based on prior experience and (implicit or explicit) past relative frequency count. Such an interpretation is only applicable to events that are essentially similar.

According to this last interpretation, the cognitive processes underlying probability judgments of unique and (in contrast) essentially similar events may be fundamentally different. In addition, it has often been claimed that feedback on correctness of responses (or judgments) is a necessary condition for obtaining good calibration (e.g., Kadane and Lichtenstein 1982), or at least for improving it (e.g., Lichtenstein and Fischhoff 1980; Oskamp 1962). According to the current account, however, this conclusion is restricted to tasks that are composed of essentially similar events. Indeed, those few studies that reported very

good calibration (e.g. Keren (1987) using expert bridge players; Keren (1988a) using a perceptual task; Tomassini et al. (1982) using accountants; Murphy and Winkler (1977) using weather forecasters), all employed events or items that can be judged as highly similar on essential dimensions.

It should be emphasized that the judgment of the extent to which events are related or essentially similar is mainly subjective and measured on a continuous scale. In practice, one would have to develop similarity measurement procedures for evaluating the degree of perceived similarity within each set of stimuli or events to be predicted. The significant commonalities shared by different events (or physical stimuli), are achieved by encoding these events into an organized system of categories. Applying the enormous amount of research on categorization (e.g., Glass and Holyoak 1986: chapter 5; Medin and Smith 1984; Smith and Medin 1981), should serve as an important input for constructing means of assessing degree of relatedness among different events and stimuli.

Calibration and independence

Both exchangeability as well as relatedness (i.e., the distinction between unique and essentially similar events) are not identical to the notion of independence. Formally, given a probability measure P on the same outcome space, events A and B are said to be independent if $P(A \cap B) = P(A) * P(B)$. Alternatively, independence can be stated in terms of conditional probabilities: event A and B are independent if $P(A) = P(A/B)$, that is if the probability of A is independent of outcome knowledge of B. Harrison (1977) notes that 'adherents to the subjective view of probability are inclined to speak of informationally *independent events*, the added modifier reflecting the view that subjectively assessed probabilities represent the assessor's state of information' (p. 321).

Independence and exchangeability are not the same: for instance, independence requires that the probability of two events both occurring, be the product of their single event probabilities, while exchangeability does not specify the nature of the joint probability as long as it is the same for all possible pairs (Kadane and Lichtenstein 1982). The main difference between the two is that independence excludes the possibility of 'learning through experience' (De Finetti 1970), thus

prohibiting the use of any frequentistic data for making inferences; exchangeability is defined such as to retain the possibility of being influenced by experience and permitting corresponding inferences.

Relatedness is fundamentally different from both exchangeability and independence. The latter two concepts are both defined solely in terms of *outcomes*, and their distributional properties. Relatedness, in contrast, is a broader concept associated with the informational value of outcome knowledge, and depending on the nature of the events (or items) under consideration. Relatedness is thus identified by cognitive processes and stimulus characteristics, such as perceived similarity of attributes underlying the relevant items or events. Hence, events can be independent and yet constitute related items. For example, consider a bridge player who has been exposed to thousands of bridge hands. The different hands are strictly independent (assuming that each hand was obtained after well shuffling the deck of cards) but nevertheless form a set of related items. General-knowledge questions, on the other hand, are both independent (the outcome of knowing the answer to one question is independent of knowing the answer to another question) and unrelated (assuming they are drawn from different contexts).

Strictly speaking, independence is defined in terms exclusively external to the assessor, and according to Harrison (1977) is thus incompatible with the interpretation of subjective probability. Harrison (1977) adds that events are never perceived as completely independent, in that knowledge about one event will tell the assessor something about his own characteristics as a numerical assessor of probabilities, and that in turn may affect the assessments of future events. For example, a so-called 'overconfident' assessor can provide very high probability estimates, but after sufficient number of trials (with feedback) may realize that generally these assessments are too high, and consequently decide to lower systematically all subsequent estimates. Similarly, Kadane and Lichtenstein (1982) proposed that independence among items in the empirical research has been rare. They add that this non-independence is usually so small that it virtually has no impact on the results. In my opinion, this assertion should be treated as an open empirical question.

In the present context it is important to distinguish between two interpretations of independence: the formal one which is exclusively based on outcomes and characteristics of the outside world, and *perceived independence* which is an individual's subjective assessment

and interpretation of whether a given set of events is dependent or not. Perceived independence is often incongruent with the formal definition: for instance, the gambler's fallacy is a situation where people fail to realize that different events (e.g., roulette spins) are independent. On the other hand, the conjunction fallacy (Tversky and Kahneman 1983) is an example where people fail to realize the dependency between two events, and which leads to incoherent probability judgments (see also Erev and Cohen (1989) for an interesting demonstration in the context of probability judgments). It is perceived independence that may affect the quality of calibration, and it is this notion (encapsulated in the concept of relatedness) that has often been neglected in calibration studies. Future studies should look into possible factors affecting perceived independence, and how they may influence the quality of calibration.

Coherence requirements

Another set of criteria proposed by Lindley et al. (1979) for evaluating probability assessments is what they termed *syntactic rules*. The essence of these criteria is to assure that 'the relations between assessments should be governed by the laws of probability' (p. 147). Yates (1982) termed it *internal consistency*, to be distinguished from *external correspondence*, which refers to the degree of correspondence between probability assessments and reality.

Most essential in this set of requirements is coherence, which is a key concept for the subjectivist viewpoint (e.g., De Finetti 1970). A set of probabilities is said to be coherent if it does not lead to a Dutch book, that is no gambles can be constructed from such a set that would yield a certain loss independent of the observed outcome (e.g., Kadane and Lichtenstein 1982).

Formal treatments (but not empirical) by advocates of the subjectivist school suggest that a person with coherent assessments is expected to be well calibrated (e.g., Dawid 1982; Kadane and Lichtenstein 1982; Lad 1984). [6] According to Lad (1984), calibration is a concept that can

[6] Kadane and Lichtenstein claim that under the condition of outcome feedback coherence alone implies calibration. Tversky (1974) suggests that De Finetti's coherence conditions could be viewed as necessary, yet may not be sufficient in a broader view that includes expressing opinions in a manner consistent with the entire web of beliefs. Wallsten and Budescu (1983) further proposed that 'in the light of the developments in nonadditive probability theory, one may not even wish to consider De Finetti's condition as necessary' (p. 105). Whether coherence indeed implies good calibration is a pure empirical (descriptive) question.

be applied only to *all* probability assessments made by a given person, and similarly coherence applies to the composite of all probabilities specified. He even goes as far as suggesting that 'when calibration is considered as a global property of the entire belief distribution, then every coherent specified assessment has been shown to be well-calibrated' (p. 220). As with calibration, it is questionable whether tests of coherence can be meaningfully applied to events that are unrelated and are thus essentially unique.

Without engaging in a lengthy discussion on coherence, it is important to point out that empirical calibration studies suggest that people's probability judgments satisfy a primary aspect of coherence: virtually all calibration curves reported in the experimental literature are strictly monotonically increasing. [7] The few exceptions in which this rule is violated can be accounted for by chance factors, resulting from an insufficient number of observations for estimating a particular point (this is another reason for the recommendation mentioned earlier, that the number of observations for estimating each point on the curve should be explicitly presented).

One possible interpretation that can be derived from the fact that empirical calibration curves are strictly monotonically increasing, is that the assessors involved are capable of discriminating between more likely and less likely events (or responses), but can express the differences only on an ordinal scale. According to this explication, people's poor calibration is just a question of scaling (Fischhoff et al. 1977). In light of such an interpretation, the validity of the so-called overconfidence phenomenon becomes questionable, and suggested explanations to account for it should be re-examined. For instance, the finding that the degree of overconfidence (as traditionally measured by formula (1)) is positively correlated with difficulty of the task implies, under this interpretation, that the more difficult the task the more difficult is the discrimination between more and less likely events, resulting in a relative flat calibration curve. According to such an account overconfidence should not necessarily be inferred from a flat calibration curve.

The proposition that subjective probabilities are expressed only up

[7] This conclusion is based on calibration curves that were constructed by lumping up all responses in intervals of 0.1 (or 10%). A different procedure that would use narrower intervals may reveal some local violations of the increasing monotonicity, but would not change the overall pattern. It will only suggest that people are unable to distinguish between very small differences in probabilities.

to an ordinal scale, would certainly not be endorsed by proponents of the subjective school. According to the view adopted in the present paper, however, this remains an open empirical question.

Are predictions and confidence ratings identical?

Several authors (e.g. Ronis and Yates 1987; Wagenaar and Keren 1986; Wright and Wishuda 1982) have proposed to distinguish between two types of probability assessments that are employed in calibration studies: one type concerns probability assessments of future events that are unknown to anyone at the time the assessments are made. Examples are predicting events such as election outcomes, rain, effect of a drug on a patient, or hitting an oil well. The second type of probabilities refers to assessing one's own knowledge, that is the probability that one's answer is correct. In that case the answer is known, although not to the assessor. Note that this distinction is also apparent in natural language: probability assessments are usually used in the context of predicting the future, whereas assessments of (past and present) knowledge are marked by confidence ratings. It is important to realize, however, that from a formal Bayesian viewpoint, all probabilities are treated alike as subjective and personal. Any alleged differences between probabilities associated with future events and confidence ratings attached to knowledge questions, refer solely to possible cognitive processes underlying probability assessments.

The major difference between the two types of assessments is based on where uncertainty is located: in the case of future events the source of uncertainty is inherent in the real world, whereas in the case of knowledge, the locus of uncertainty is in the assessor himself. Howell (1971; Howell and Burnett 1978) made a related distinction between probability assessments where uncertainty is *internal* or skill derived, and *external* or environment derived. He applied the terms respectively to events that can or cannot be controlled, and presented some experimental evidence suggesting that overconfidence occurs mainly in situations with internal uncertainty where the assessor may have some (apparent or real) control.

Kahneman and Tversky (1982) also employ the same terms, but the distinction they made is closer to the one adopted here. Specifically, they propose to distinguish between uncertainty attributed to the external world where probabilities reflect the relative strength of differ-

ent competing dispositions (which they further classify into distributional or singular modes, as mentioned above), and uncertainty attributed to internal states of knowledge. (Smith, Benson and Curely (1991) have questioned the usefulness of the internal–external distinction, because 'uncertainty, like knowledge, presumes both knower and known', and one cannot exist without the other. While the statement is obviously correct, it is not clear why it undermines the importance of the internal–external distinction as an important characteristic of uncertainty.)

The processes underlying probabilistic assessments of future compared with past or present events may differ in several ways. First, it is possible that when one judges external uncertainties, the uncertainty, or probability factor, is an integral part of the conceptual representation, or mental model, of the real world. As such, it may be readily available for introspection and modification. When internal uncertainties are concerned as in estimating correctness of general-knowledge questions, assessors are asked for their knowledge about knowledge, i.e., metacognition. The uncertainty that is being measured may be an assessment of the adequacy of the mental model itself, that is monitoring the processes leading to the model and judging its reliability.

Another difference may be implied from Howell's (1971) suggestion that overconfidence in cases of internal uncertainty stems from the assessors' perceived control. Recent experimental evidence, however, does not necessarily support the assertion that internal uncertainty is always accompanied with overconfidence. Keren (1988a), using a within-subjects design, obtained the usual overconfidence for general-knowledge questions but not for a simple perceptual task (which is also characterized by internal uncertainty). It is possible that perceived control is especially pronounced in tasks that require intellectual capacities (e.g., Dawes 1980). Wright and Wishuda (1982) suggested that there may be a positive social utility attached to expressing certainty when knowledge and intelligence are concerned.

Given these alleged differences between assessing future and past (in particular, general-knowledge questions) events, the question whether such differences also exist in practice remains an empirical query. Wright and Wishuda (1982) compared a general-knowledge task with questions regarding future events. Overconfidence was significantly higher for the knowledge task, but unfortunately task and knowledge were confounded: the general-knowledge questions were much more

difficult compared with prediction of future events (60% vs. 82%). In a subsequent study, Wright (1982) controlled for difficulty and obtained slightly more confidence in the general questions task (though the problem in this study was that both tasks were relatively easy). Fischhoff and MacGregor (1982), on the other hand, failed to find any differences between confidence assessments regarding forecasts and those pertaining general-knowledge questions. Finally, Ronis and Yates (1987) report significantly higher levels of overconfidence for general-knowledge questions compared with predictions of basketball outcomes. The only finding common to all the above studies was the greater reluctance to express complete certainty about future events. To summarize, the current experimental literature suggest that:

(1) Despite the robust finding of overconfidence with general-knowledge questions (at a sufficient level of difficulty), this finding cannot be generalized to all kinds of past or present events (e.g., perceptual tasks). (2) There is some support for the assertion that people tend to be somewhat less confident when predicting future events, but the evidence is not conclusive. The major question of interest, for which we are far from having a conclusive answer, is whether the cognitive mechanisms underlying prediction of future and past events are indeed different. In a recent paper, Wolford et al. (1990) demonstrated that the conjunction fallacy (Tversky and Kahneman 1983) is more likely to occur with known rather than unknown outcomes, and proposed that different processes may guide probability assessments for these two types of events. Though the studies by Wolford et al. were not in the context of calibration, they support the conjecture that, from a descriptive viewpoint, probability assessments of present and future events may be fundamentally different.

Methodological issues in eliciting subjective probabilities

Ideally, probability judgments (and corresponding calibration curves) should be independent of the method by which they were elicited; in practice, however, this is highly questionable. There is sufficient evidence to suggest that probability assessments are not always invariant across different elicitation procedures. In the present section the possible effects of procedural variables and elicitation methods on quality of calibration are critically examined.

Instructions and task

A plausible explanation for the observed overconfidence and miscalibration exhibited by many people is that they do not understand the task, misinterpret it, or, more specifically, do not comprehend the (probability) response scale. A natural remedy would be to provide people with elaborate and explicit instructions. Lichtenstein and Fischhoff (1981) presented subjects with detailed instructions that had little effect. They concluded that the sources of miscalibration and overconfidence are deeply rooted, and result from subtle cognitive difficulties that cannot be resolved by lengthy explanations.

Another potential source for improving calibration is by training. For instance, Lichtenstein and Fischhoff (1980) attempted to improve calibration and reduce overconfidence by providing their subjects with extensive training with feedback. The results showed a modest improvement, most of which was achieved at the early stage of training. Arkes, Christensen et al. (1987) also attempted to reduce overconfidence by using a specific training method: they exposed one of their groups to some practice questions that presumably appeared to be easy, but in fact were quite difficult (as was evident from the low percent of correct answers). Subjects in this group, who also received feedback after the practice trials, were much better calibrated (in fact, they showed some underconfidence) compared with other control groups.

Why does extensive training yield only modest improvements, and hardly generalizes to other tasks? One simple answer has to do with the nature of the stimuli, namely the degree to which they constitute related items. Indeed, most of the studies attempting to improve calibration by training used general-knowledge questions which, as was claimed, constitute unrelated or unique items. Moreover, most training studies employed 'mechanical' manipulations (Fischhoff 1982), or what Keren (1988b) termed 'procedural' methods. Such methods do not yield better probability judgments and improved calibration – they at best result in some technical corrections. For instance, subjects who continuously receive feedback that their probabilities are too high will naturally lower their probabilities across the line. The question is whether this change by itself implies that they have now become better assessors? The most disturbing finding obtained from training studies is that whatever modest improvement is achieved, it hardly ever generalizes to other tasks.

More promising training procedures should attempt to restructure the task (Fischhoff 1982; Keren 1988b). Procedural aspects may interact with cognitive processes underlying probability assessments, and thus investigating procedural effects may provide some initial cues for a more comprehensive theory. For instance, Koriat et al. (1980), used a procedure forcing their subjects to list pro and con reasons (for each two-alternative question) before choosing an answer and assessing the corresponding probability. This procedure produced a marked improvement that led the authors to suggest that overconfidence may be caused by a bias to justify the chosen alternative by favoring confirming evidence (Einhorn and Hogarth 1978) and disregarding evidence that is inconsistent with the chosen answer. If overconfidence can indeed be accounted for by the confirmation bias, then different framings of the same questions may lead to different levels of overconfidence. The procedure employed by Koriat et al. may have altered some cognitive processes involved in probability judgments. Unfortunately, the extent to which the effect of this procedure generalizes to other tasks has not been tested.

Number of alternatives and choice vs. no choice procedures

The number of alternatives (as in multiple-choice tests) presented to the assessor is important for two reasons. First, as mentioned above, it determines the nature of the probability scale and the corresponding chance level associated with it. A second consideration is entailed from the objectives of obtaining the probability forecasts (or confidence ratings), and can be best illustrated in the context of knowledge tests with multiple choice items.

Consider a test item with k alternatives, implying that the assessor is limited to k different responses (or values). Shuford and Brown (1975) noted that if the person's state of knowledge and degree of uncertainty regarding the particular question can actually assume more than k different responses, it would be impossible to have each different response associated with a single state of knowledge. The assessor would therefore have to use the same response for more than one state of knowledge. Consequently, the restricted response set would act as a filter inserted in the communication channel between the assessor and the user of the assessments. This limitation can only be resolved by increasing the number of response options. Since different people

possess different states of knowledge (and the same person may possess different states of knowledge at different times) and these are unknown, Shuford and Brown (1975) recommend using a large number of response options, thus providing the opportunity of transmitting more information.

In practice it may be difficult to determine the most appropriate number of options, and the choice of the specific alternatives will always be subjective. Also, a too large number of options may have several disadvantages, beside the ones mentioned above. In particular, comparing and assessing the likelihood of many different alternatives in a systematic manner may become impossible due to limited memory and processing capacity. Some empirical evidence exists, suggesting that the number of alternatives may systematically affect probabilistic assessments. Teigen (1983) found that the majority of his subjects gave estimates that added up to unity (i.e., 100%) only in the two-alternative case. As the number of alternatives increased, the probability totals increased in direct proportion to the number of alternatives. These results led Teigen (1983) to propose that, at least in his study, subjects adopted a non-distributional conception of probability. Apparently, this supra-additivity (i.e., the tendency to overestimate probabilities of mutually exclusive outcomes of the same event so that they add up to more than 100%) is also extended to verbal probabilistic expressions (Brun and Teigen 1988).

Indirect evidence concerning possible effects of number of alternatives is also obtained from a study by Fischhoff et al. (1978) on fault trees. They presented subjects with a varying number of reasons (presented in the form of a fault tree) for the event 'a car fails to start'. Subjects' estimated failure probabilities for a particular reason varied widely depending on the number of alternative reasons presented.

The number of alternatives also determine the extent to which the assessor has to make a choice decision. Ronis and Yates (1987) pointed out the difference between 'no choice' situations in which the assessment is made of a predefined target event (e.g., 'it will rain tomorrow', or 'the operation will succeed'), and between choice situations commonly used in the laboratory in which the assessor has first to choose between k alternatives, and then assign a probability to the chosen event or response (a two-stage process). Beside the difference in scales that are used in these situations (0 to 100% under no choice conditions; $100/k$ to 100% under k choice conditions), Ronis and Yates (1987)

proposed that the act of choosing an answer before assigning a probability may affect the judgment process and consequently the probability assigned. They referred to Bem's (1967) self-perception theory suggesting that the act of choice increases the attractiveness of the chosen alternative and the commitment to it (and in turn, partially account for the bias of searching for confirming evidence). Consequently, they predicted higher confidence in the choice condition.

To test their hypothesis, Ronis and Yates used three conditions: a usual binary 'choice' condition with a scale of 0.50 to 1.0, a 'no choice' condition with a 0 to 1.0 scale and, to avoid the confounding of choice with type of scale, a third condition of 'choice' combined with a scale of 0 to 1.0. The results were inconclusive, and in any event did not support the choice hypothesis. A possible reason for the failure of the choice hypothesis is that in the 'no choice', condition, attention is immediately directed to the given target event or response, and hence other possible events or responses are neglected thus leading to overconfidence.

In their final conclusions, Ronis and Yates (1987) recommended the choice procedure with a scale of 0.50 to 1.0 and rejected the no choice procedure because it produced high overconfidence. They also rejected the choice procedure with a full scale of 0 to 1.0 since subjects in this condition often used probabilities below 0.50, i.e., below chance level, for their preferred answers.

Probabilities below chance level

The empirical observation of subjects assessing probabilities below chance level deserves further consideration. One possibility is that these subjects simply misunderstand the fundamental concept of a probability scale in which case they should be simply removed from the sample (generally, for all calibration studies, it would be desirable to have a screening procedure that would exclude such subjects from the sample).

There is, however, an alternative view to explain such behavior. Recently, I conducted a calibration study (unpublished) using general-knowledge questions and employing a two-alternative procedure. Subjects received explicit instructions that only confidence ratings between 50% and 100% were admissible, and all followed the instructions (with an exception of one subject, whose data were eliminated from the sample). As the task was sufficiently difficult (approximately 66%

correct responses), we obtained the usual phenomenon of overconfidence. At the end of the task, subjects were asked to estimate the number of items (out of 50) they believed to have answered correctly. Comparing this aggregate measure with the actual number of correct responses, overconfidence vanished almost completely (it was not statistically significant). In other words, mean confidence in the aggregate estimates was significantly lower than mean confidence over the individual items. There was, however, a problem in interpreting the data: in the initial task of assessing confidence for each individual item, subjects were explicitly instructed to avoid estimates that are below chance level. In the subsequent task of estimating the aggregate, no boundaries were specified. Apparently, more than 14% of the subjects gave an aggregate estimate of less than 25 items, which is below chance level. When these subjects were excluded from the sample, mean overconfidence in the aggregate condition obviously increased, and was not significantly different (although still lower) from the mean overconfidence of the individual items.

Two conclusions are to be drawn from the above study: one is that when the appropriate comparison is used there is no difference between the amount of confidence assigned to individual items and to the aggregate. The second conclusion is more subtle: the fact that subjects assign estimates below chance level, suggests that they either do not understand the concept or at least are not aware of it in the present context. This conclusion is not necessarily restricted to those subjects whose estimates were below chance level. It is probably applicable to some other subjects as well, except that it was not observed in the present study. If these subjects were exposed to a more difficult task, it cannot be ruled out that at least some of them would have also used below chance level estimates.

The assignment of probabilities below chance level may also apply to individual items, except that in the latter case subjects are explicitly told to avoid such estimates, so they are not observable. Experimenters usually assume, especially in the case of general-knowledge questions, that in case of complete ignorance subjects will toss a mental coin to decide which alternative to choose, and then assign a probability of 0.5 to the chosen alternative.

The notion of tossing a mental coin, however, entails several problems: first, tossing a mental coin implies that the response is randomly chosen, and previous research has shown that people are very poor in

producing random sequences (Wagenaar 1972). Second, and more important, most people would feel reluctant to employ a random generator for answering general-knowledge questions (knowledge is antithetical to randomness). There is ample evidence suggesting that when faced with lack of knowledge (especially in the case of general-knowledge questions), subjects will revert to use inferences (e.g., Allwood and Montgomery 1987; Keren 1987; May 1986). Many subjects fail to realize that their inferences are error prone and do not discount their probability assessments accordingly, which consequently leads to overconfidence (Keren 1987). There are, however, large individual differences (further discussed later in this article), and some subjects may be aware that their inferences could frequently yield wrong answers. Normatively, if we assume that subjects never guess randomly yet often rely on inferential solutions, then assessing the number of correct responses to be below chance level should not necessarily be considered as inappropriate. However, estimates below chance level are permissible only when aggregate estimates are required, but not with estimates of individual items. In the latter case there is a certain asymmetry: those subjects who fail to realize that their inference may be wrong exhibit overconfidence. On the other hand, subjects who want to discount for the possibility of an erroneous inference are restricted by a lower boundary (i.e., chance level).

There is another aspect to the same problem. Many general-knowledge tasks contain some items that can be termed 'deceptive' or 'misleading'. Such items are characterized by a percentage of correct responses that is significantly below chance level. For example, in a study by Keren (1988a), subjects were asked which country has a larger population: Israel or Nepal? 87% of the subjects believed that Israel has a larger population and assigned a mean confidence of 0.70 for responding correct. In fact, the population of Nepal is almost three times as large as that of Israel. Supposedly, since Israel is continually mentioned in the news, subjects may have considered this fact as a useful cue which apparently in this particular case it was not.

The major conclusion, from this and similar examples, is not that subjects are using fundamentally wrong inferences. In fact, it is most probable that a positive correlation exists between country size and media exposure. Subjects, however, do not realize that the inference is probabilistic (and more general, that it is error prone), and fail to discount their confidence ratings accordingly. More important for the

present context are two related conclusions: one is that even if subjects do not possess the relevant knowledge they are not likely to toss a mental coin, and their mean response may well lie below the chance level. Second, the existence of misleading items creates an experimental bias for producing overconfidence, since even on items that are scored significantly below chance level, subjects are not allowed to provide confidence ratings below 50%.

The so-called 'misleading items' are not restricted to general-knowledge questions tasks. Wagenaar and Keren (1986) report an eyewitness study in which subjects were shown slides presenting a car–pedestrian accident. Later, subjects were presented with picture pairs and asked which one they have seen before and how certain they were in their choice. On 5 out of the 15 test trials accuracy was less then 50%, and the scores for the two worst items were 18% and 21%. Evidently, when subjects failed to remember they did not toss a mental coin. As Wagenaar and Keren proposed, 'the subjects were not guessing; they thought that there was a good reason to pick the wrong answer' (p. 90).

Response mode

Different elicitation procedures and response modes may be judged by a formal analysis to be equivalent, yet in practice may yield substantially different results (Hershey et al. 1982). Empirical demonstrations show that subjective probability distributions assessed by different elicitation techniques and response modes often produce results that are different and may even be inconsistent with each other (e.g., Schaefer and Borcherding 1973; Seaver et al. 1978; Stael van Holstein 1971).

The effect of response mode is important because it may shape in one way or another the cognitive processes underlying probability judgments, and consequently the quality of calibration. A central question in this regard is whether the numbers that represent subjective probabilities exist in the assessor's (or subject's) internal representation independent of the method being used to elicit these numbers, or are constructed in the elicitation process itself? This is an intricate question for which there is no simple answer (the question is further discussed in the concluding comments). If the relevant uncertainty is already part of the internal representation of the event, then response mode may affect the stage of mapping, or translating, these uncertainties to a particular

scale. If assessment of uncertainty is part of the elicitation process itself, then response mode may affect the assessment of uncertainty at the stage at which it is formed.

A particularly interesting distinction between two classes of elicitation procedures, which is directly linked with alleged different cognitive processes, has been recently suggested by Peterson and Pitz (1988). They distinguish between two general prototypes of assessment tasks. One, the most commonly used in psychological experiments on calibration (and exclusively in studies with general-knowledge questions), in which the assessor has to choose between n alternatives ($n = 1,2,3,\ldots m$) and then state the probability that the chosen alternative is correct. In the other procedure, the assessor has to state values of an uncertain quantity (e.g., the population of France is?) that are associated with two (or more) predetermined fractiles of the distribution (e.g., 0.05 and 0.95). Note that these two procedures are usually associated with discrete and continuous propositions, respectively. Peterson and Pitz (1988) pointed out that in the former task the outcome is fixed (the chosen answer to each question) and assessors are requested to provide a probability judgment for the outcome. In contrast, in the second task the probability is fixed (i.e., 0.05 and 0.95) and assessors are asked to give a range (that is variable) of answers as their response.

Peterson and Pitz (1988) proposed that the first type of task measures the person's belief that a previous stated prediction is correct, and termed it *confidence*. In contrast, the second task (using fractiles) is measuring the person's beliefs about the variability of possible outcomes (i.e., variability of a subjective probability distribution) which they termed *uncertainty*. This last term can be interpreted in the context of information theory. Indeed, the authors provide empirical demonstrations that what they termed confidence and uncertainty are affected in different ways by the available information. In particular, 'uncertainty is determined by the number of different predictions that can be generated, whereas confidence is influenced by salient factors that people believe affect the accuracy of their predictions' (p. 85).

Even within each of the two types of assessments, different elicitation procedures can be used that may yield different results. Phillips and Edwards (1966), employing discrete propositions, have shown that probability assessments varied depending on whether they were directly estimated on a scale from zero to one, compared with the use of 'odds' or 'log odds'. With continuous variables, Seaver et al. (1978) studied

five different procedures for assessing subjective probability distributions over continuous variables that varied in response mode (probabilities, odds, and odds on a logarithmic scale), and type of response (uncertainty measure, or a value of an unknown quantity). The results showed that the fractile procedures were inferior to the procedures requiring odds or probabilities as a response. There is also evidence that the order by which the fractiles are solicited may affect the results (Selvidge 1980).

Irrespective of the particular mode of elicitation, all the methods discussed above relate to numerical responses. Uncertainty, however, can also be expressed in verbal terms, and one may even claim that such a response mode is more natural and easier to understand (Wallsten et al. 1986). Erev and Cohen (1990) report that their subjects had a preference to express and convey probabilistic judgments in the verbal mode, and at the same time when serving as decision makers preferred to receive probabilistic information in numerical terms. They termed this finding as the 'communication mode preference' (CMP) paradox. One possible way to account for the CMP is in terms of costs and benefits: constructing numerical probability judgments may require more cognitive effort and entail a higher level of commitment which may be the reason why assessors prefer to formulate their uncertainties in a verbal form. Formally, it may seem that numerical judgments are more efficient and precise, which may explain why receivers of information prefer the numerical mode. Apparently, empirical results do not support this intuition unequivocally: whereas Budescu et al. (1988) report that expressions of numerical probabilities are more efficient, Erev and Cohen fail to find such an advantage and proposed that the difference may be due to the nature of the experimental situation under study.

Several researchers tested whether a consistent mapping exists from the domain of verbal terms to the domain of numerical responses (e.g., Beyth-Marom 1982; Brun and Teigen 1988; Budescu and Wallsten 1985; Lichtenstein and Newman 1967). The general finding from all these studies is a modest agreement at the group level, yet large intersubject variability exists in the numerical values attached to different probability terms. The interpretation of different verbal terms is also highly vulnerable to context effects (e.g., Brun and Teigen 1988).

The ambiguity associated with probability terms should not necessarily, and solely, be attributed to intrinsic characteristics of verbal

expressions. The possibility that sources of vagueness are inherent in any expression of uncertainty, verbal or numerical, is not ruled out. Both modes of response can be equally vulnerable to potential biases and misinterpretations. For example, Teigen (1988) has shown that both numerical and verbal probabilities tend to be overestimated in connection with multi-outcome events.

The main normative advantage of numerical responses is that they lend themselves naturally (though not without difficulties) to assessment and evaluation. Thus, despite the various interpretations and difficulties involved, we can still speak meaningfully of calibration studies of numerical probability estimates. Whether verbal probabilities are well calibrated is much more difficult to assess, since they have to be transformed for that purpose into numerical values. [8]

Notwithstanding, there are two (related) reasons for the interest in studying verbal probability expressions. One, mentioned already, is that at least some of the ambiguities associated with verbal probabilities are equally applicable to the numerical mode, that is, they share a common source of ambiguity. For example, Fox (1986) proposed that the terms *probable* and *plausible* are sometimes being used interchangeably (whereas plausibility is expected to be supported by arguments, probability should be supported by direct or indirect empirical evidence). Whether an assessor will adopt one or the other interpretation may depend on different factors (such as context), but the same potential confusion applies equally to both numerical and verbal probabilities. More generally, both numerical and verbal probabilities are equally vulnerable to framing and context effects.

A second reason for the interest in studying verbal probabilities is related to the cognitive mechanisms underlying probability assessments. If indeed the initial natural representation of uncertainty is in the verbal mode then, even in those circumstances where a numerical response is explicitly required (as in calibration studies), subjects may

[8] To assess calibration of verbal judgments, the experimenter has to transform these expressions into numerical terms. Wallsten et al. (1986) describe such a procedure. However, such procedures are somewhat hampered by individual differences that exist in subject's interpretation of verbal probabilities. Which are more reliable, numerical probabitities provided by assessors (supposedly after they have translated their vague feelings regarding uncertainty), or numerical terms obtained by the experimenter after transforming assessors verbal expressions, is currently left as an open question.

first initiate a verbal assessment that may later be transformed into a numerical response.

Loss functions and goals

The discussion up to now has centered on factors that are external to the assessor. Obviously, these factors may interact with the cognitive processes evoked by probability judgments and in turn affect calibration. In the present section, we discuss internal factors of the assessor, such as loss functions, strategies, motivation and individual differences.

Loss functions and scoring rules

Two related problems are essential for properly interpreting probability assessments in general, and calibration studies in particular. One is the 'honesty' question that draws directly upon the elicitation procedure: since subjective probability assessments exist solely in the assessor's mind (Murphy and Winkler 1970), there is no way to determine whether they agree with the reported probabilities. How can one control, or at least encourage, assessors to produce assessments that accurately reflect their judgment or internal state? The second question concerns the evaluation procedure (Friedman 1983): According to what criteria should the quality of probability judgments be evaluated? Probability assessments may be largely influenced by the loss function adopted by the assessor, that is by the rewards and penalties associated with each combination of response and outcome, and that should be taken into account in the evaluation process. Levi (1985) pointed out that the squared error loss function underlying the Brier score (which is the most common evaluation measure), may not be optimal for different decision makers. Different assessors may hold different attitudes and preferences regarding the type and magnitude of errors they are willing to tolerate, and this may further differ from one context to another. For a proper evaluation process it is important to ensure that the evaluator and the assessor (or in the laboratory, experimenter and subject) assume the same loss function.

Both issues of eliciting the 'true' probabilities and ensuring that the proper loss function is integrated in the evaluation process are supposed to be resolved by the use of *scoring rules* (e.g., Murphy and

Winkler 1970). A scoring rule is a function that assigns a score to every possible combination of a probability assessment and the actual value (or correct response) associated with it. *Proper scoring rules* are constructed such that assessors can maximize their expected score if, and only if, they set their overt probability response equal to their 'true' internal assessment (Winkler and Murphy 1968).

A large body of literature exists offering formal analyses of different scoring rules and their properties (e.g., Friedman 1983; Murphy 1972a 1972b, 1973; Murphy and Winkler 1970; Nau 1985; Winkler 1969). There are several practical problems associated with the introduction of scoring rules:

(1) A scoring rule assumes the existence of one 'true' underlying probability distribution in the assessor's mind. Whether such a single distribution really exists, and whether the assessor is always aware of it is highly questionable (Hogarth 1975).

(2) A scoring rule can be effective only to the extent that: (a) the assessors understand exactly how their probability statements are evaluated by the scoring rule, and (b) the assessors are making an attempt to follow and maximize this scoring rule (Friedman 1983). The first assumption may often not hold, especially when complex scoring rules are concerned. With regard to the second assumption, as it is impossible to validate that assessors are reporting their 'true' subjective probability, it is similarly impossible to validate whether they are indeed employing (and correctly) a given scoring rule.

(3) As pointed out by Murphy and Winkler (1970), although all scoring rules are supposed to encourage 'honesty', some scoring rules may be more likely to encourage honesty than others. This is a natural question for psychological investigation.

(4) The extent to which a scoring rule may encourage careful assessment may depend on the nature of the rule. Murphy and Winkler (1970) suggest that sharper scoring rules are more sensitive since deviations from optimality are more costly. Von Winterfeldt and Edwards (1982) convincingly argue that most scoring rules, at least in the experimental laboratory, suffer from the flat maxima phenomenon implying relatively small differences in payoffs for optimal and non-optimal decisions. How sensitive assessors are with regard to different scoring rules has not yet been established empirically; in any event, researchers are strongly advised to take account of the potential effects

of the flat maxima phenomenon in the process of designing and interpreting experiments (for further detail, see von Winterfeldt and Edwards 1982).

(5) A scoring rule is a translation of certain goals to be achieved, and thus the assumption is that such goals exist and are well defined. In reality, this assumption may often be invalid. Moreover, frequently there are several goals to be achieved, and if two or more of these goals are conflicting it may be difficult, if not impossible, to transform them into a coherent scoring rule.

With few exceptions, little empirical research has been conducted to investigate the effectiveness of scoring rules. Jensen and Peterson (1973) compared the three most popular rules (log, quadratic, and spherical) and found little differences in the probabilities inferred from each of these three rules. However, probabilities became less extreme with increased steepness in the functions relating score to assessed probability.

In another study, Fischer (1982) made a direct attempt to evaluate the impact of scoring rules. Based on four different cues, Fischer asked his subjects to predict grade point average (GPA), for several hypothetical freshman students, by assigning probabilities to one of four possible intervals (of GPA). He employed a truncated logarithmic scoring rule that is characterized by 'flatness' for moderate and large values of the probability assigned to the true value, and drops sharply for values lower than 0.25.

The major effect of the scoring rule was to deter subjects from using very low probabilities due to the potential heavy penalties associated with such probabilities. No other statistically significant effect was evident though, compared with the control groups, the scoring-rule groups were both less confident and closer to the predictions of a Bayesian classification model (see Fischer (1982) for details).

The effect of the scoring rule in Fischer's study was certainly limited. However, no generalizations to other tasks and other scoring rules would be appropriate. Under certain real-life circumstances, assessors may be extremely sensitive to the nature of the loss function (e.g., physicians). Clearly, a more substantial empirical research program is needed. Surprisingly, most psychologists have neglected the issue, and virtually all calibration studies that were conducted in the laboratory failed to use an explicit scoring rule. Customarily, these experiments

have been analyzed by using the Brier score which is a quadratic scoring rule. Whether subjects in these studies assumed the same rule is highly questionable (Levi 1985). In fact, one major possible source for the large individual differences observed in calibration studies may simply be the adoption of different strategies and different scoring rules by different subjects.

Unlike laboratory investigations, real-life situations often carry with them natural scoring rules as for example in medicine: under most circumstances, physicians adopt a payoff matrix that assigns a greater cost to a false negative diagnosis than to a false positive one (e.g., Scheff 1963). Fryback and Erdman (1979) warn, however, that all the results in the medical field were obtained under somewhat artificial conditions (where primary attention was given to diagnosis), and question whether these results can be generalized to the real world.

Although natural loss functions are often ill-defined, and can therefore not be expressed in a strict formal way, they may nevertheless present key guidelines to the assessor. The literature on scoring rules is dominated by a normative perspective. An important question from a descriptive viewpoint, and one that has been completely ignored, concerns the natural scoring rules adopted by subjects when such a rule is not given by an external authority. Self-developed scoring rules, though not precisely formulated, may have a larger impact on the assessor's probability judgments compared with artificial scoring rules, and may be less susceptible to the flat maxima phenomenon.

Motivational and social factors

Given that scoring rules have rarely been used, and their moderate impact on those cases in which they were used, the question remains whether assessors construct an internal scoring rule which guides their probability judgments. Although it is unlikely that assessors would adopt by themselves one of the formal scoring rules, there are at least two important psychological factors that may guide their behavior and may account for the pervasive overconfidence reported by so many studies.

Dawes (1980) suggested that we overestimate the power of our intellectual abilities, a tendency that has been reinforced by the fast pace of technological developments. This overestimation may apply not just to general knowledge questions, but to any task that relies on

intellectual capacities. Related to this argument is the claim that because intellectual capacities and knowledge are highly regarded, it is desirable that confidence in own knowledge should be demonstrated. More generally, overconfidence may in particular be pronounced under conditions in which the assessor holds some (apparent or real) control on the task.

Another factor that may enhance probability judgments causing an apparent overconfidence, is own involvement in outcomes coupled with wishful thinking. A persuasive demonstration in an ecologically valid setting, is reported by Babad (1987). He asked a large sample of football spectators to predict outcomes, and has shown that the stronger a person felt affiliated with a team, the more likely he or she was to assign a win to that team. Even predictions made at half time, when the favorite team trailed decisively, were characterized by a pervasive tendency of wishful thinking. A similar observation was reported by Keren (1987): he noticed that amateur bridge players assigned higher probabilities to the success of their own contracts than to contracts of their opponents.

Finally, overconfidence may simply result from some social norms in which the assessor responds to social expectations. A distinct example is the case of physicians: most patients expect their physicians to be confident about their diagnosis (whatever it is), and interpret uncertain statements (even when they are normatively justified) as reflecting the physician's poor expertise. The physician's response to such expectations may be concealed in inflated probabilities. Obviously, the expectations from experts to be confident in their field of expertise is not limited to the medical field.

Numerical measures: The analysis of calibration studies

Several numerical measures exist, most of them in the form of scoring rules, to evaluate probabilistic assessments and analyze calibration studies. It is important to understand the assumptions and underlying logic of these measures, since the particular measure used may strongly affect the interpretation of calibration studies.

In virtually all calibration studies, the data are arranged by grouping similar assessments, usually within ranges (e.g. all assessments between 0.50 and 0.59, etc.), and analyzed over subjects as well as items. Estes

(1964) presented an insightful discussion on the dangers involved in using group data in the study of probability learning. He concluded that 'on the one hand incautious inferences from averaged data may lead us to attribute to individuals relationships that exist only in groups; on the other hand, failure to make judicious use of statistical analysis may result in our failing to discover relationships which do characterize the individual but which become apparent only in appropriately averaged data' (p. 93). This warning is equally applicable to analyses of calibration studies. Ideally, any complete evaluation of calibration studies should include a separate analysis of the individual assessors, and a separate analysis of the individual items.

Possible individual differences

Averaging over items (but not over individuals) should provide insight into the nature and characteristics of the assessors. Assessors may differ in either their substantive expertise (i.e., the specific domain knowledge on which the assessments are made) or in an alleged skill of producing appropriate probabilistic assessments. Lichtenstein and Fischhoff (1977), using a general knowledge task, analyzed separately the calibration of subjects who had a higher or lower substantive skill (based on percentage correct responses). Overall, the confidence ratings of the 'better' subjects (i.e., subjects with a higher percentage of correct responses) was higher than the confidence ratings of subjects with a lower performance. However, the quality of calibration (as measured by the Brier score) of the two groups was equally poor.

Ronis and Yates (1987) looked into individual consistency with respect to calibration across two rather different tasks. They found moderate consistency in mean confidence ratings, but only weak consistency in all other components of judgment quality. They concluded that probability judgment is not a unitary trait, and that different skills and knowledge are required from different tasks.

A study by Keren (1987) showed remarkable differences in calibration depending on substantive skill: asked to forecast the outcomes of different bridge games, expert players were remarkably better calibrated compared with amateur bridge players (though the latter had a long history of experience). As noted, there is indeed some evidence that experts may often be well calibrated, but generalizations in this respect are unwarranted. Moreover, good calibration (when it exists) is ap-

parently restricted to the domain of expertise and does not transfer to other domains (Keren 1985).

A more fundamental question concerns possible individual differences in being well calibrated (in the normative sense). Several difficulties are involved in investigating this question: first, one has to ensure that the different assessors are equal in their substantive knowledge. Just relying on performance based on percent correct predictions (or responses) may not be sufficient, especially when the number of items is relatively small (which is usually the case). Second, if the analysis is to be performed separately on each assessor, a large number of observations from each individual is necessary if the results are to be reliable. Obtaining subjects for a large number of observations entails practical difficulties. In particular, subjects may get bored and impartial after a certain number of assessments, and high incentives may be needed to keep them motivated. Finally, once such differences are observed, one has to search for other attributes that are correlated with these differences.

Some indication for possible individual differences in the normative sense have been reported by Lichtenstein and Fischhoff (1980). Using a general-knowledge task, they found that a third of their subjects were well calibrated prior to any training. Wright and Phillips (1976) report some weak relationships between personality measures and verbal expressions of uncertainty. Currently, however, no conclusive statements can be made regarding such individual differences and their nature.

Item difficulty

The data from any calibration study can be split and arranged by items (averaging over subjects). Such an analysis is useful for several purposes: first, it enables one to detect misleading or deceptive items, and the researcher may want to delete those items for further analyses. Second, it may provide the variance of the 'difficulty' scale. As items are more homogenous with regard to difficulty, the smaller should be the range or variance of the corresponding probability assessments. This last statement assumes that the measure of difficulty applies equally to all assessors, that is that the group of assessors is homogenous with regard to substantive knowledge, an assumption that is not necessarily correct.

Regarding difficulty, it has often been claimed (e.g., Lichtenstein et al. 1982) that difficulty is highly correlated with over- or underconfidence. As the task becomes more difficult, overconfidence increases and quality of calibration decreases. Whether this finding has any psychological significance is highly questionable. Consider, for instance, a two-alternative task of extreme difficulty on which the assessor cannot get more than 50% correct predictions (except by chance). If the assessor is explicitly instructed not to use probabilities below 50%, then by definition, such an assessor can never be underconfident: he or she can only be overconfident (or perfectly calibrated). A similar argument (though less forceful) would apply to difficult tasks in which only 55% or 60% of the predictions are correct. Now consider the opposite situation in which the task is extremely easy, and the assessor has all predictions correct without any exception. This time, by definition, the assessor can never be overconfident but only underconfident (or perfectly calibrated). Again, a similar argument can be stretched to other easy tasks in which performance is somewhat below 100%. The above analysis suggests a built-in mechanism (without any psychological substance) that would enhance overestimation of difficult items, and underestimation of easy items. Note that virtually all calibration curves reported in the literature show underconfidence for low probabilities (i.e., 0.50 and 0.60) and overconfidence for high probabilities close to 1.00 (see for example fig. 1).

An alternative explanation (not excluding the previous one) for the interaction between difficulty and over/underconfidence, albeit one that applies only to laboratory experiments, is in terms of subjects' expectations. According to this account, the laboratory setting creates an expectation of an intermediate level of difficulty. The two extremes, namely a task that is either so difficult that performance is on a chance level or a task that is so easy that performance will always be perfect or close to perfect, are assumed unlikely. Consequently, subjects may anchor on a probability estimate that would reflect intermediate difficulty, like 75% in two-alternative tasks. Whenever an item is perceived to be very easy or very difficult, subjects would adjust accordingly but as is well known such adjustments are usually not sufficient (e.g., Tversky and Kahneman 1974) and thus would lead to under or overconfidence respectively.

It should also be noted, that a more difficult task is not necessarily associated with larger overconfidence. For example, Keren (1988a)

employed a perceptual and a general-knowledge task for which the mean proportion of correct responses was 0.67 and 0.71 respectively. Thus, despite the fact that the perceptual task was more 'difficult', it did not yield overconfidence as did the 'easier' task of general-knowledge questions.

With few exceptions, most calibration studies have not analyzed separately assessors and items. One reason for avoiding such analyses is often related to the limited number of observations that can be obtained in practice. Notwithstanding, researchers are advised to carry out such analyses, not as a substitute but rather as a complement to traditional analyses.

Measures of calibration

The Brier score

The Brier score is a proper scoring rule (Brier 1950; Lichtenstein and Fischhoff 1980; Murphy 1972a, 1972b, 1973) and the most often used in analyzing calibration studies with discrete propositions. For purposes of simplicity and ease of exposition, the following discussion is limited to 'two alternatives' situations but is readily generalizable to conditions with multiple alternatives. Given N items, a probability assessment p_i, and a corresponding variable c_i which is 1 when the correct prediction or response have been chosen and otherwise 0, the Brier score is a squared error loss function on predictions and outcomes defined as

$$BR = 1/N \sum_{i-1}^{N} (p_i - c_i)^2. \tag{2}$$

Given further that probability assessments are arranged in t probability ranges or categories, and that n_t is the number of observations in category t where $\times n_t = N$, the Brier score is given as

$$BR = 1/N \sum_{t=1}^{T} \sum_{i=1}^{n_t} (p_t - c_{it})^2. \tag{3}$$

Ideally, the Brier score should be 0 implying a perfect correspondence between assessments and reality. The larger the Brier score, the larger

the discrepancies between the assessments and the real occurrence of the events.

The Brier score can be partitioned and decomposed in several ways. Such decompositions highlight the fact that 'good' probability assessments have more than one facet, enable a more refined analysis of a set of forecasts, and can point out more accurately possible weaknesses of the assessor. A commonly used decomposition proposed by Murphy (1972a, 1972b, 1973), is

$$BR = C(1 - C) + 1/N \sum_{t=1}^{T} n_t(p_t - C_t)^2 - 1/N \sum_{t=1}^{T} n_t(C_t - C)^2, \qquad (4)$$

where C is the overall proportion of correct predictions, and C_t is the proportion of correct predictions in range t $C_t = 1/n_t \sum_{i=1}^{n_t} C_{it}$.

The first term is simply an *outcome index*: it indicates the proportion of predictions borne out. In the case of general-knowledge questions it provides an index for the proportion of correct responses and thus has occasionally been referred to as 'knowledge' (Lichtenstein and Fischhoff 1977). In a certain respect the term marks the 'difficulty' of the task, and thus serves as a reference point for interpreting the other two components. Difficulty, however, is a relative term, and a more proper interpretation of this term would be 'expertise' (i.e., substantive proficiency in the task, not in calibration).

The second term has been referred to by Murphy (1973) as *reliability* and by Lichtenstein and Fischhoff (1977, 1980) as *calibration*. It measures the goodness of fit between the probability assessments and the corresponding proportions of correct responses (or the deviation of the calibration curve from the 45 deg line).

Finally, the third term, called *resolution*, is a measure of the variance of the probability assessments. It is subtracted from the first two terms, and thus the larger it is the better (smaller) the Brier score. Resolution is an indirect measure of the information contained in a set of predictions, or alternatively measures the assessor's ability to discriminate between the likelihood of different events, and thus is an indication of the forecaster's skill. If this term is 0 (i.e., the forecaster is repeatedly providing the same probability estimate) the assessor is said to have 'no skill'; yet, as Yates (1982) pointed out, the assessor can still be perfectly calibrated (if the calibration term is also 0). To assess the appropriate (constant) p still requires a considerable 'baseline knowledge'.

The interpretation of a resolution score is further complicated by the fact that the score is not independent of the mean proportion correct predictions (or responses). Sharp et al. (1988) correctly observed that as an assessor's predictions improve, the total variance in the distribution of correct and incorrect answers – which is the upper bound for the resolution score – decreases. Sharp et al. (1988) suggest that a more appropriate assessment of resolution is the ratio of resolution to the outcome index (i.e., the ratio between the third and the first term in (4)). This ratio is in essence equivalent to the commonly used effect size measure, eta^2.

Several additional points should be made regarding Murphy's decomposition. First, in contrast to the first term, calibration and resolution are under the assessor's control and in that respect measure his or her skill. [9] Second, as shown by Yates (1982), calibration and resolution are not completely independent of each other. Third, and not withstanding, there is a sort of internal conflict between being well calibrated (i.e., minimizing the discrepancy between probabalistic estimates and the corresponding reality) and achieving a high resolution (i.e., maximizing the amount of information transmitted). The Brier score contains implicit instructions that may be incongruent (i.e., minimize calibration and at the same time maximize resolution). [10] There is no normative guideline that would dictate how to reconcile the two opposing goals, and different assessors may develop different strategies depending on what they want to achieve on each of these two dimensions.

A further problem with the partitioning of the Brier score stems from the (empirical) fact that variations in the first component tend to be much larger than the variations in calibration or resolution (Lichtenstein and Fischhoff 1980). The total Brier score is therefore mainly determined by the first term that reflects external factors, and is relatively insensitive to variations in the last two components that are

[9] There is a tendency to interpret these two terms as measuring what Winkler and Murphy (1968) have termed the normative skill, namely the expertise of generating appropriate probabilistic assessments. For obtaining low calibration and high resolution, however, substantive skill is required as well. The decomposition, therefore, cannot separate between these two types of expertise that are apparently deeply interrelated.

[10] This resembles the internal conflict often contained in the instructions used in experiments on attention and information processing in which subjects are required to respond as accurate as possible and at the same time as fast as possible. Such conflicting instructions may be confusing (Edwards 1961) and open the room for different strategies that the subject can use.

supposedly under the assessor's control. In other words, substantive knowledge is the most dominant component of the Brier score. One should thus be cautious in interpreting empirical results in terms of the three components.

Finally, one should be careful in computing significance tests since the sampling properties of the Brier score (or its components) are unknown. Fischhoff and MacGregor (1983) proposed to use the jackknife procedure (Mosteller and Tukey 1977), which is suitable when the sampling distribution is unknown or when computed on unstable estimates (as mentioned, the number of observations in a particular probability range can occasionally be small thus yielding an unstable estimate).

Yates (1982) developed an alternative decomposition which was later extended to multiple-event situations (Yates 1988), and can be applied to either discrete or continuous forecasts. Yates called Murphy's reliability (the second term) *reliability-in-the-small*, and has shown that it can further be broken down into:

$$\text{Reliability-in-the-small} = S_p^2 + (P - C)^2 - 2S_{pc} + \text{Murphy's resolution}, \tag{5}$$

and substituting it in eq. (4) yields

$$BR = C(1 - C) + S_p^2 + (P - C)^2 - 2S_{pc}. \tag{6}$$

The term S_p^2 is the variance of the assessor's forecasts, and should be minimized (so that BR is minimized). Note that according to Murphy's partitioning the variance of the forecasts (as reflected in the resolution term) should be as large as possible, whereas according to Yates' (1982) decomposition the variance should be as small as possible. Yates was aware of this contradiction, and qualified his advice by suggesting that S_p^2 should be minimized *given* the assessor's fundamental forecasting abilities as represented by the covariance S_{pc}. But what these fundamental forecasting abilities are is one of the key questions to be answered by calibration research, and Yates remains silent on this issue. This problem is another reflection of the internal antagonism between accuracy (as measured by calibration) and information (as measured by resolution).

The term $(P-C)^2$ (P standing for mean probability assessment over all items) is termed by Yates *reliability-in-the-large*, and is supposed to reflect the assessor's ability to match the mean forecast (over the entire collection of items) to relative frequencies. The interpretation of this term raises again several essential problems that were discussed earlier: do reliability-in-the-small and in-the-large measure different attributes? Empirically, do these measures yield different results, and if so how could it be explained? Is there more justification to use a frequentistic criterion for reliability-in-the-large?

What can be learned from the two partitions (and other possible ones) described above? Yates (1982) and Ronis and Yates (1987) claim that decompositions of the Brier score could be potentially valuable tools in the study of basic cognitive processes underlying probability judgments. The different components may measure different skills that are supposedly required for producing adequate probability judgments. Unfortunately, we do not yet know what these skills are (or should be). Moreover, because of the inconsistencies described above, in particular regarding calibration and resolution, the danger exists that looking at various components may lead to unwarranted different (and sometimes incongruent) conclusions.

An interesting and stimulating paper recently published by Murphy and Winkler (1987), offers a somewhat new outlook on the assessment and verification of probabilistic forecasts. In particular, Murphy and Winkler propose a framework which is based on the joint distribution of forecasts and observations which, they claim, contains all the relevant information. Denoting the probabilistic forecasts by f and the corresponding observations by x, the joint distribution can be represented as $p(f, x)$. Murphy and Winkler (1987) further show that the joint distribution can be further represented in two conditional forms:

$$p(f, x) = p(x \mid f)p(f). \tag{7a}$$

$$p(f, x) = p(f \mid x)p(x). \tag{7b}$$

The two terms in eq. (7a) describe two characteristics: $p(x \mid f)$ relates to calibration whereas $p(f)$ is the marginal distribution of the forecasts and, according to Murphy and Winkler (1987), measures refinement indicating how often different probability values are being used. Conceptually, resolution and refinement pertain to the same attribute.

The more novel part of Murphy and Winkler's framework is con-

tained in eq. (7b). Here, $p(f \mid x)$ provides the likelihood associated with different forecasts given the observation x. Murphy and Winkler (1987) interpret these likelihoods as indicating the extent to which a forecast *discriminates* among different observations (values of x). Finally, $p(x)$ gives the probability for different values of x (the base-rate) and thus describes the forecasting situation rather than the forecasts or the forecaster.

An essential difference between the two equations is that in (7a) the emphasis is on 'the labels assigned to the forecasts and toward the actual properties of the forecasts when they are taken at face value' (Murphy and Winkler 1987: 1334). In other words, the nature of the probability scale is of utter importance. In contrast, in (7b) the forecasts can be perfectly discriminating regardless of the particular scale of f. We may thus be talking about two different attributes (that supposedly correspond to different cognitive processes) namely, (a) the possibility to discriminate between different events, and (b) the ability to transform these differences into an appropriate probability scale. Indeed, several authors (Ferrell and MvGoey 1980; Keren 1987; Koriat et al. 1980;) proposed a two-stage model in which the first stage involves searching one's knowledge and attempting to detect the proper response, and a second stage where the evidence is reviewed and confidence in the chosen alternative is translated into a numerical probability. A proper application of Murphy and Winkler's framework may assist in testing the validity of the two-stage model. Note that the proposed framework was developed in the context of weather forecasting, and is thus applicable only to situations with essentially similar or related events.

Other calibration measures

Murphy and Winkler (1970) discuss several scoring rules beside the Brier rule. They distinguish between situations under which the assessor's utility function for the score is linear (such as the Brier score) or nonlinear. If the assessor's utility function is linear then maximizing the expected score is equivalent to maximization of expected utility. This is not necessarily the case when the utility function is nonlinear. Daan (1981) offers an exhaustive review of possible measures for evaluating probabilistic assessment (verification scores). We do not elaborate on these rules, since most of them are complicated, non transparent, and at best can serve a selected group of experts.

Signal detection measures

Several investigators (e.g., Ferrel and McGoey 1980; Levi 1985; Smith and Ferrel 1983) have noticed the similarity between signal detection and calibration tasks. A calibration task can be described as a task that requires a discrimination between true and false statements, and correspondingly a probability assessment (a rating response in the signal detection terminology). Ferrel and McGoey (1980) developed a mathematical model, based on signal detection theory, to describe the calibration of discrete subjective probabilities. The model was applied to data reported by Lichtenstein and Fischhoff (1977) and resulted in adequate fit. Unfortunately, the model provides little insight into the possible cognitive processes governing probability assessments. As the authors themselves noticed 'it is thus more a model of the task than of the respondent's behavior' (p. 52). Similar conclusions have been reached by Smith and Ferrel (1983).

Glenberg and Epstein (1985) employed the so-called 'Confidence-Judgment Accuracy Quotient' (CAQ) for evaluating calibration in a comprehension task. The measure is defined as

$$CAQ = \frac{\text{Mean Confidence (correct)-Mean Confidence (wrong)}}{\sqrt{\text{Pooled variance of confidence correct and wrong}}}, \qquad (8)$$

and is analogous to d' in signal detection analysis.

The apparent similarity between calibration and signal detection should be treated cautiously. A fundamental difference between the two is that in traditional signal detection studies the distributions of signal and noise are based on a *stimulus classification*, whereas in calibration studies the corresponding classification is made in terms of *subjects' responses*. Generally, the focus of signal detection is mainly on discriminability, which is only part of the probability assessment process. Measures like d', or similar measures like the CAQ, may at best serve as additional estimates for resolution, but are not adequate measures for calibration.

General discussion

Despite the increasing interest and the expanding number of articles on calibration, some essential conceptual issues remain unresolved,

which consequently hinder further progress. There are, in my opinion, two main (and related) reasons for that state of affairs: first, most researchers adopted a rigid normative view based on a formal mathematical model, and were unable to bridge between the normative framework and empirically obtained descriptive data. Implicitly or explicitly, researchers have assumed that the mind is processing (or should process) probabilistic information in accordance with strict rules derived from the formal model. Whether in fact the cognitive system is operating according to such mental rules, analogous to those of the formal system, is highly questionable. Similar doubts regarding the isomorphism between the logical deductive formal system and a corresponding alleged cognitive setup, have been recently raised by Rips (1990).

A second problem hampering the study of how the cognitive system processes probabilistic information concerns the multiple definitions of probability. [11] As was shown above, calibration studies have adopted a hybrid probabilistic framework in which the logical, frequentistic, and subjective interpretation of probability have been assembled in an often ambiguous way, which cannot be reconciled. The consequences are that even if we disregard the first problem, namely the doubtful similarity between the normative and descriptive domains, there is a lack of an unambiguous acceptable yardstick to which observed probabilistic assessments can be compared.

Employing a distinction that has been recently proposed by Rips (1990) [12], I propose to classify calibration studies according to two perspectives: the *strict* view, and the *loose* view. The strict view relies rigorously on the formal (normative) theory. It assumes that the process underlying probability judgments is a discrete step by step process, analogous and parallel to the logic underlying the formal theory. According to this view, even people without formal training have an intuition for making probability judgments (that admittedly is error prone) which is well structured and resembles the formal system. Consequently, probability assessments in this approach are appraised by precise and rigorous criteria derived from the formal model.

[11] In this regard the similarity with the deductive reasoning breaks, since there is little disagreement among researchers concerning the interpretation of the formal deductive system.

[12] Rips (1990) proposed this distinction in the context of reviewing the literature on reasoning. Rips' ideas are applied here to the specific context of calibration and probability assessments with some necessary modifications.

The loose view is less formal and assumes a more fluid and adaptable system. According to this view, the processes underlying probability judgments are continuous and interactive (rather than sequential), and consist of mutual adjustments in the cognitive system resulting in a strength of belief that is then translated into a subjective probability. Unlike the strict outlook which is deductive in nature, the loose view is founded on an inductive mode of reasoning, and validation is based on 'inductive strength' that is not entirely based on a logical form. A stimulating model of probability judgments, compatible with the loose view, has been recently proposed by Smith et al. (1991).

The strict and the loose view differ in several respects on each of the major issues reviewed in this article, consequently resulting in different perspectives and interpretations of calibration studies. On the conceptual level, the strict view applies probability theory to an abstract world (the logical interpretation), or to the physical world (the frequentistic interpretation). Uncertainty in this approach is interpreted as a construct that is external to the human assessor, and where the human's assessor role is ('technically') to map these external uncertainties into numerical or verbal statements according to predetermined rules. Performance is then compared with an 'ideal assessor' analogous to the 'ideal receiver' represented by a mathematical simulation of a physical system designed to behave optimally according to Signal Detection Theory (Coombs et al. 1970).

In contrast to the strict view, the loose approach assumes that uncertainty is an internal attribute of the assessor in which the cognitive system plays a central role. Given that humans' information processing system is selective and limited in capacity, probability assessments are made through a continuous process of adopting to the partial current information available at any point in time. Note that although the loose approach is in some respects closer to the subjectivist (or Bayesian) viewpoint, they are by no means identical. A fundamental difference between the strict and the loose view is that the former is based on a *phenomenal* approach whereas the latter is mainly *epistemological* in nature.

The differences between the strict and the loose view are apparent in the conceptualization and interpretation of probability assessments, as well as in their evaluation. Consider the incompatibility between elicited subjective probabilities and their assessment by a frequentistic criterion. Following the strict view, this incompatibility (unless completely

ignored) is supposed to be resolved *within* the formal abstract interpretation of probability theory. For instance, exchangeability is a reconciliation attempt offered within the strict view. The important point to realize is that exchangeability is mainly a phenomenal construct which is defined by the properties of the stimulus distribution. Implicitly, it is assumed that the cognitive system can identify exchangeable events and utilize the concept in the probability assessment process, but no empirical evidence exists suggesting that this notion has any implications for the cognitive system. Adherents of the Bayesian approach have postulated that exchangeability is entirely a subjective notion, but it has never been explicated in what respect.

The concepts of relatedness and essential similarity, introduced earlier as alternatives to exchangeability, are more in line with the loose approach since they are grounded on psychological attributes such as categorization and similarity judgments, and are mainly epistemological constructs. Admittedly, little research has been conducted on either of these two concepts within the domain of probability judgments.

A similar difference between the two views concerns the notion of independence: the strict view adopts statistical independence (which is determined by formal abstract characteristics of the model), whereas *perceived* independence is a characteristic of the cognitive system in accordance with the loose view.

The two views also differ in the manner by which probability assessments are appraised. The measures of calibration that were briefly reviewed above, all stem from the strict approach. Despite their formal character, these measures are not without problems. For example, how should one weight the relative importance of calibration and resolution, and how should the internal conflict between these two terms be resolved? Because the loose view is centered on the assessor rather than on an abstract formal model, the evaluation criteria employed in this perspective may not always be as rigorous (and exclusive) as that of the strict view. For instance, Smith et al. (1991) suggest that truth values of certain propositions cannot be always determined in which case a justification criterion may often be the only operative criterion. Protocol analysis, as employed by Allwood and Montgomery (1987), may be a helpful tool in this respect.

The two views differ also with regard to methodology and elicitation methods that were discussed in the second part. According to a rigid interpretation of the strict view, probability judgments (as made by the

'ideal' assessor) should be invariant with regard to the method by which they have been elicited. Thus issues like number of alternatives, choice vs. no-choice procedures, framing, and response mode are irrelevant. According to the loose view which assumes a dynamic cognitive system (that despite its adaptability to changing conditions, is characterized by limited memory and processing capacity), these factors are an inherent part of the cognitive system and should therefore be incorporated in the conceptual framework. More generally, probability judgments should be assessed in the broader context in which they were elicited. In addition, unlike the strict view, the loose view accepts individual differences.

To obtain reliable and coherent assessments, the strict view offers the use of formal proper scoring rules. It disregards the difficulties involved in employing scoring rules such as complexity, the flat maxima phenomenon, and other problems elaborated on earlier. In the loose view those difficulties are recognized and, moreover, the existence of natural (and less formal) scoring rules is not ruled out. In fact, one of the goals according to the loose view should be to identify the characteristics of such natural rules, which may also be affected by different motivational factors, as were mentioned earlier.

In practice, the two views are often intertwined and the distinction between the two approaches is better described along a continuum. In my view, it is often the confounding of the two views that has hindered the development of a more theoretical descriptive framework within the loose view. Note that the adoption of the loose perspective does not imply the abandoning of a formal interpretation of probability theory. However, it implies a less rigid reliance on the formal model and a broader perspective. For instance, following the strict view most empirical studies assume, explicitly or implicitly, that cognitive processes invoked by probability judgments are the same, regardless of the stimulus material and the nature of the task that is being used. As should be apparent from the present review, it is highly questionable whether the processes underlying probabilistic assessments of general-knowledge questions are the same as those underlying probabilistic statements of weather forecasting. The loose view accepts that there is more than one descriptive model to characterize the processes underlying probability assessments. This is, for example, the reason for the proposal to distinguish between related and unrelated items, or be-

tween probability assessments of single as compared to multiple (repeated) events.

The strict view has never resolved completely the issue regarding multiple interpretations (e.g., logical, frequentistic, subjective) and as suggested, ended up with a hybrid model which is a major source of many of the problems discussed in this paper. According to the loose view, which interpretation is the most appropriate one should depend on the particular task and stimulus material that are being used. Thus, the loose view denies the necessity for a single 'correct' interpretation. Consequently, according to the loose view, there is not one single set of criteria for evaluating probability assessments. For example, whereas good reasons may exist for evaluating weather forecasts by some frequentistic measures, the interpretation of such measures when applied to general-knowledge questions is highly controversial.

Indeed, one may doubt whether calibration is a meaningful measure for all possible tasks. Basically, calibration is supposed to measure the accuracy of probability assessments, but the question still remains, 'accuracy in what sense'? The strict view conceives the 'true' probability to be reflected by relative frequencies measures. But is there indeed a 'true' probability? Phillips (1970), in accordance with the loose view, correctly pointed out that such a 'true' or 'objective' probability often does not exist, and that a probability cannot be right or wrong. Calibration then is at best one possible way to assess probability judgments, and apparently only under certain circumstances.

A central problem with the strict view is indeed its strict definition of calibration, namely the accuracy by which probability judgments correspond to reality. The loose approach is based on a broader standard which may be termed as the adequacy of probability judgments (and in which calibration is just one of a larger criteria ensemble). What are adequate probabilities? Since any probability statement is meant to convey information, it should be accurate as far as possible. However, as I have argued, the criterion for accuracy when applied to probability judgments is often ambiguous and controversial, and under certain circumstances meaningless. Moreover, the information contained in a probability statement should be evaluated not just by precision, but also by amount and quality, as for instance offered by the measure of resolution.

Both calibration and resolution (which stem from the strict view) allude to final outcomes. It should be remembered, however, that a

probability judgment constitutes a decision, and most researchers suggest (implicitly or explicitly) that the judgment regarding the quality of a decision should be based not just on outcome, but also on the inferred process by which it has be reached (e.g., Keren 1986; Simon 1978; Wright and Murphy 1984; Vlek et al. 1984). Judgment by *process* is equally applicable to probability assessments, and should comprise an essential part of the loose outlook. Smith et al. (1991), have recently claimed that since subjective probabilities are and will remain subjective, an understanding of the cognitive processes underlying such judgments is essential. They present a cognitive analysis of subjective probability judgments, and discuss a belief-processing model in which 'reasoning is used to translate data into conclusions, while judgmental processes qualify those conclusions with degrees of belief'. The paper by Smith et al. may signal the begininng of a shift toward the loose view.

References

Allwood, C.M. and H. Montgomery, 1987. Response selection strategies and realism of confidence judgments. Organizational Behavior and Human Decision Processes 39, 365–383.

Babad, E., 1987. Wishful thinking and objectivity among sport fans. Social Behavior: An International Journal of Applied Social Psychology 4, 231–240.

Baron, J., 1988. Thinking and deciding. Cambridge: Cambridge University Press.

Bem, D.J., 1967. Self-perception: An alternative explanation of cognitive dissonance phenomena. Psychological Review 74, 183–200.

Beyth-Marom, R., 1982. How probable is probable? A numerical translation of verbal probability expressions. Journal of Forecasting 3, 473–474.

Brun, W. and K.H. Teigen, 1988. Verbal probabilities: Ambiguous, context dependent, or both? Organizational Behavior and Human Decision Processes 41, 390–404.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review 78, 1–3.

Budescu, D.V. and T.S. Wallsten, 1985. Consistency in interpretation of probabilistic phrases. Organizational Behavior and Human Decision Processes 36, 392–495.

Budescu, D.V., S. Weinberg and T.S. Wallsten, 1988. Decisions based on numerically and verbally expressed uncertainties. Journal of Experimental Psychology: Human Perception and Performance 14, 281–294.

Chan, S., 1982. Expert judgments under uncertainty: Some evidence and suggestions. Social Science Quarterly 63, 428–444.

Christensen-Szalanski, J.J.J. and J.B. Bushyhead, 1981. Physicians' use of probabilistic information in a real clinical setting. Journal of Experimental Psychology: Human Perception and Performance 7, 928–935.

Coombs, C.H., R.M. Dawes and A. Tversky, 1970. Mathematical psychology. Englewood Cliffs, NJ: Prentice-Hall.

Daan, H., 1981. Scoring rules in forecast verification. Technical report 81–10, Royal Netherlands Meteorological Institute, De Bilt.

Dawes, R.M., 1980. 'Confidence in intellectual judgments vs. confidence in perceptual judgments'. In: E.D. Lantermann and H. Feger (eds.), Similarity and choice. Bern: Hans Huber.

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

De Finetti, B., 1970. Theory of probability, Vols. 1, 2. New York: Wiley.

Edwards, W., 1961. Costs and payoffs are instructions. Psychological Review 68, 275–284.

Einhorn, H.J. and R.M. Hogarth, 1978. Confidence in judgment: Persistence in the illusion of validity. Psychological Review 85, 395–416.

Erev, I. and B.L. Cohen, 1990. Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. Organizational Behavior and Human Decision Processes 45, 1–18.

Estes, W.K., 1964. 'Probability learning'. In: A.W. Melton (ed.), Categories of human learning. New York: Academic Press.

Ferrel, W.R. and P.J. McGoey, 1980. A model of calibration for subjective probabilities. Organizational Behavior and Human Performance 26(4), 32–53.

Fischer, G.W., 1982. Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting. Organizational Performance and Human Performance 30, 352–369.

Fischhoff, B., 1982. 'Debiasing'. In: D. Kahneman, P. Slovic and A. Tversky (eds.), Judgement under uncertainty: Heuristics and biases. Hillsdale, NJ: Erlbaum.

Fischhoff, B., S. Lichtenstein and P. Slovic, 1977. Knowing with certainty: The appropriateness of extreme confidence. Journal of Experimental Psychology: Human Perception and Performance 3, 552–564.

Fischhoff, B. and D. MacGregor, 1982. Subjective confidence in forecasts. Journal of Forecasting 1, 155–172.

Fischhoff, B. and D. MacGregor, 1983. Categorical confidence. Unpublished report, Decision Research, Eugene, OR.

Fischhoff, B., P. Slovic and S. Lichtenstein, 1978. Fault trees: Sensitivity of estimated failure probability to problem representation. Journal of Experimental Psychology: Human Perception and Performance 2, 330–334.

Fox, J., 1986, 'Knowledge, decision making, and uncertainty'. In: W.A. Gale (ed.), Artificial intelligence and statistics. New York: Addison-Wesley.

Friedman, D., 1983. Effective scoring rules for probabilistic forecasts. Management Science 29, 447–454.

Fryback, D.G. and H. Erdman, 1979. Prospects for calibrating physicians' probabilistic judgments: Design of a feedback system. Proceedings IEEE International Conference on Cybernetics and Society, 340–345.

Glass, A.L. and K.J. Holyoak, 1986. Cognition. New York: Random House.

Glenberg, A.M. and W. Epstein, 1985. Calibration of comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition 11, 702–718.

Hacking, I., 1975. The emergence of probability. Cambridge: Cambridge University Press.

Hampton, J.M., P.G. Moore and H. Thomas, 1973. Subjective probability and its measurement. Journal of the Royal Statistical Society 136A, 21–42.

Harrison, J.M., 1977. Independence and calibration in decision analysis. Management Science 24, 320–328.

Hershey, J.C., H.C. Kunreuther and P.J.H. Schoemaker, 1982. Sources of bias in assessment procedures for utility functions. Management Science 28, 936–954.

Hogarth, R., 1975. Cognitive processes and the assessment of subjective probability distributions. Journal of the American Statistical Association 70, 271–289.

Howell, W.C., 1971. Uncertainty from internal and external sources: A clear case of overconfidence. Journal of Experimental Psychology 89, 240–243.

Howell, W.C. and S.A. Burnett, 1978. Uncertainty measurement: A cognitive taxonomy. Organizational Behavior and Human Performance 22, 45–68.

Jensen, F.A. and C.R. Peterson, 1973. Psychological effects of proper scoring rules. Organizational Behavior and Human Performance 9, 307–317.

Kadane, J.B. and S. Lichtenstein, 1982. A subjectivist view of calibration. Decision Research Report 82-6.

Kahneman, D. and A. Tversky, 1982. Variants of uncertainty. Cognition 11, 143–157.

Keren, G., 1985. On the calibration of experts and lay-people. Paper presented at the 10th conference on Subjective Probability, Utility and Decision Making, Helsinki.

Keren, G., 1986. On the judgment and measurement of 'good' decisions. IZF report 1986-31, Soesterberg.

Keren, G., 1987. Facing uncertainty in the game of bridge: A calibration study. Organizational Behavior and Human Decision Processes 39, 98–114.

Keren, G., 1988a. On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. Acta Psychologica 67, 95–119.

Keren, G., 1988b. 'Cognitive aids and debiasing methods: Can cognitive pills cure cognitive ills'. In: J.P Caverni, J.M. Fabre and M. Gonzalez (eds), Cognitive biases. Amsterdam: North-Holland.

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

Kyburg, H.E. Jr., 1968. Bets and beliefs. American Philosophical Quarterly 5, 54–63.

Kyburg, H.E. Jr., 1983. Rational belief. The Behavioral and Brain Sciences 6, 231–273.

Kyburg, H.E. and H.E. Smoekler, 1964, Studies in subjective probability. New York: Wiley.

Lad, F., 1984. The calibration question. British Journal of Philosophy of Science 35, 213–321.

Levi, K., 1985. A signal detection framework for the evaluation of probabilistic forecasts. Organizational Behavior and Human Decision Processes 36, 143–166.

Lichtenstein, S. and B. Fischhoff, 1977. Do those who know more also know more about how much they know? Organizational Behavior and Human Performance 20, 159–183.

Lichtenstein, S. and B. Fischhoff, 1980. Training for calibration. Organizational Behavior and Human Performance 26, 149–171.

Lichtenstein, S. and B. Fischhoff, 1981, The effects of gender and instructions on calibration. Decision Research technical report PTR-1092-81-7, Eugene, OR.

Lichtenstein, S., B. Fischhoff and L.D. Phillips, 1982. 'Calibration of probabilities: The state of the art to 1980'. In: D. Kahneman, P. Slovic and A. Tversky (eds.), Judgment under uncertainty: Heuristics and biases. Hillsdale, NJ: Erlbaum.

Lichtenstein, S. and J.R. Newman, 1967. Empirical scaling of common verbal phrases associated with numerical probabilities. Psychonomic Science 9, 563–564.

Lindley, D.V., A. Tversky and R.V. Brown, 1979. On the reconciliation of probability assessments. Journal of the Royal Statistical Society 142A, 146–180.

Lusted, L.B., 1977. Study of the efficacy of diagnostic radiologic procedures: Final report on diagnostic efficacy. Chicago, IL: Efficacy study committee of the American College of Radiology.

Manger, T. and K.H. Teigen, 1988. The time horizon in students' predictions of grades. Scandinavian Journal of Educational Research 32, 77–91.

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 and G. Seron (eds.), New directions in research on decision making. Amsterdam: North-Holland.

Medin, D.L. and E.E. Smith, 1984. Concepts and concept formation. Annual Review of Psychology 35, 113–138.

von Mises, R., 1957. Probability, statistics, and truth. London: Allen and Unwin.
Mosteller, F. and J.W. Tukey, 1977. Data analysis and regression. Reading, MA: Addison-Wesley.
Murphy, A.H., 1972a. Scalar and vector partitions of the probability score: Part I. Two state situation. Journal of Applied Meteorology 11, 273–282.
Murphy, A.H., 1972b. Scalar and vector partitions of the probability score: Part II. *N*-state situations. Journal of Applied Meteorology 11, 1183–1192.
Murphy, A.H., 1973. New vector partition of the probability score. Journal of Applied Meteorology 12, 595–600.
Murphy, A.H. and H. Daan, 1984. Impacts of feedback and experience on the quality of subjective probability forecasts: Comparison of results from the first and second years of the Zierikzee experiment. Monthly Weather Review 112, 413–423.
Murphy, A.H. and R.L. Winkler, 1970. Scoring rules in probability assessments and evaluation. Acta Psychologica 34, 273–286.
Murphy, A.H. and R.L. Winkler, 1977, Can weather forecasters formulate reliable probability forecasts of precipitation and temparture? National Weather Digest 2, 2–9.
Murphy, A.H. and R.L. Winkler, 1987. A general framework for forecast verification. Monthly Weather Review 115, 1330–1338.
Nau, R.F., 1985. Should scoring rules be 'effective'? Management Science 31, 527–535.
Oskamp, S. 1962. The relationship of clinical experience and training methods to several criteria of clinical prediction. Psychological Monographs 76, (28, whole No. 547).
Peterson, D.K. and G.F. Pitz, 1988. Confidence, uncertainty and the use of information. Journal of Experimental Psychology: Learning Memory and Cognition 14, 85–92.
Phillips, L.D., 1970. 'The "true probability" problem'. In: G. de Zeeuw, C.A.J. Vlek and W.A. Wagenaar (eds.), Subjective probability: Theory, experiments, applications. Amsterdam: North-Holland.
Phillips, L.D. and W. Edwards, 1966. Conservatism in a simple probability inference task. Journal of Experimental Psychology 72, 346–357.
Pitz, G.F., 1974. 'Subjective probability distributions for imperfectly known quantities'. In: L.W. Gregg (ed.), Knowledge and cognition. Hillsdale, NJ: Erlbaum.
Rips, L.J., 1990. Reasoning. Annual Review of Psychology 41, 321–353.
Ronis, D.I. and J.F. Yates, 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.
Salmon, W., 1967. The foundations of scientific inference. Pittsburgh: University of Pittsburgh Press.
Schaefer, R.E. and K. Borcherding, 1973. The assessment of subjective probability distributions: A training experiment. Acta Psychologica 37, 117–129.
Scheff, T.J., 1963. Decision rules and type of error and their consequences in medical diagnosis. Behavioral Science 8, 97–107.
Seaver, D.A., D. von Winterfeldt and W. Edwards, 1978. Eliciting subjective probability distributions on continuous variables. Organizational Behavior and Human Performance 21, 379–391.
Selvidge, J., 1980. Assessing the extremes of probability distributions by the fractile method. Decision Sciences 11, 493–502.
Shafer, G. and A. Tversky, 1985. Languages and designs for probability judgment. Cognitive Science 9, 309–339.
Sharp, G.L., B.L. Cutler and S.D. Penrod, 1988. Performance feedback improves the resolution of confidence judgments. Organizational Behavior and Human Decision Processes 42, 271–283.
Shuford, E. and T.A. Brown, 1975. Elicitation of personal probabilities and their assessment. Instructional Science 4, 137–188.
Sieber, J.E., 1974. Effects of decision importance on ability to generate warranted subjective uncertainty. Journal of Personality and Social Psychology 30, 688–694.

Simon, H.A., 1978. Rationality as process and product of thought. American Economic Review 68, 1–16.

Smith, E.E. and D.L. Medin 1981. Categories and concepts. Cambridge, MA: Harvard University Press.

Smith, G.F., P.G. Benson and S.P. Curley, 1991. Belief, knowledge, and uncertainty: A cognitive perspective on subjective probability. Organizational Behavior and Human Decision Processes 48, 169–192.

Smith, M. and W.R. Ferrel, 1983. 'The effect of base rate on calibration of subjective probability for true-false questions: Model and experiment'. In: P. Humphreys, O. Svenson and A. Vari (eds.), Analyzing and aiding decision processes. Amsterdam: North-Holland.

Spetzler, C.S. and C.A. Stael von Holstein, 1975. Probability encoding in decision analysis. Management Science 22, 340–358.

Stael von Holstein, C.A.S., 1971. Two techniques for assessment of subjective probability distributions – An experimental study. Acta Psychologica 35, 478–494.

Teigen, K.H., 1983. Studies in subjective probability III: The unimportance of alternatives. Scandinavian Journal of Psychology 24, 97–105.

Tomassini, L.A., I. Solomon, M.B. Romney and J.L. Krogstad, 1982. Calibration of auditors' probabilistic judgments: Some empirical evidence. Organizational Behavior and Human Performance 30, 391–406.

Tversky, A., 1974. Assessing uncertainty. Journal of the Royal Statistical Society 36B, 148–159.

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

Tversky, A. and D. Kahneman, 1986. Rational choice and the framing of decisions. The Journal of Business 59, 251–278.

Tversky, A., S. Sattath and P. Slovic, 1988. Contingent weighting in judgment and choice. Psychological Review 95, 371–384.

Vlek, C.A.J., W. Edwards, I. Kiss, G. Majone and M. Toda, 1984. What constitutes a good decision. Acta Psychologica 56, 5–27.

Wagenaar, W.A., 1972. Generation of random sequences by human subjects: A critical survey of the literature. Psychological Bulletin 77, 65–72.

Wagenaar, W.A. and G. Keren, 1986. 'Does the expert know? The reliability of predictions and confidence ratings of experts'. In: E. Hollnagel and D. Woods (eds.), Intelligent decision aids in process environments. Berlin: Springer-Verlag.

Wallach, M.A., 1958. On psychological similarity. Psychological Review 65, 103–116.

Wallsten, T.S. and D.V. Budescu, 1983. Encoding subjective probabilities: A psychological and psychometric review. Management Science 29, 151–171.

Wallsten, T.S., D.V. Budescu, A. Rapaport, R. Zwick and B. Forsyth, 1986. Measuring the vague meanings of probability terms. Journal of Experimental Psychology: General 115, 348–365.

Winkler, R.L., 1967. The quantification of judgment: Some methodological suggestions. Journal of the American Statistical Association 62, 1105–1120.

Winkler, R.L., 1969. Scoring rules and the evaluation of probability assessors. Journal of the American Statistical Association 64, 1073–1078.

Winkler, R.L. and A.H. Murphy, 1968. Good probability assessors. Journal of Applied Meteorology 7, 751–758.

von Winterfeldt, D. and W. Edwards, 1982. Costs and payoffs in perceptual research. Psychological Bulletin 91, 609–622.

Wolford, G., H.A. Taylor and J.R. Beck, 1990. The conjunction fallacy? Memory and Cognition 18, 47–53.

Wright, G., 1982. Changes in the realism and distribution of probability assessments as a function of question type. Acta Psychologica 52, 165–174.

Wright, G. and L. Phillips, 1976. Personality and probabilistic thinking: An experimental study. Technical report 76-3, Brunel Institute of Organizational and Social Studies. Uxbridge.

Wright, G. and A. Wishuda, 1982. Distribution of probability assessments for almanac and future events questions. Scandinavian Journal of Psychology 23, 219–224.

Wright, J.C. and G.L. Murphy, 1984. The utility of theories in intuitive statistics: The robustness of theory-based judgments. Journal of Experimental Psychology: General 113, 301–322.

Yates, J.F., 1982. External correspondence: Decomposition of the mean probability score. Organizational Behavior and Human Performance 30, 132–156.

Yates, J.F., 1988. Analyzing the accuracy of probability judgments for multiple events: An extension of the covariance decomposition. Organizational Behavior and Human Decision Processes 41, 281–299.