Chapter 18

# The Calibration of Subjective Probabilities: Theories and Models 1980–94

Alastair G.R. McClelland and Fergus Bolger
*University College London*

## 18.1 INTRODUCTION

Why are individuals so often badly calibrated when making subjective probability judgements? In particular, why is overconfidence so frequently observed (the "overconfidence effect") and why does the degree of miscalibration seem to vary systematically with task difficulty (the "hard–easy" effect)? In the conclusion to their well-known review of research up to 1980 on the calibration of probabilities, Lichtenstein, Fischhoff & Phillips (1982, page 333) noted that

> ... a striking aspect of much of the literature reviewed here is its "dust-bowl empiricism". Psychological theory is often absent, either as motivation for research or as an explanation of the results.

The aim of this chapter is to review the current situation, by providing a critical analysis of a number of substantive theories and models of subjective probability judgement for discrete propositions which have appeared in the last 14 years. Little will be said regarding the empirical research which has been

*Subjective Probability*. Edited by G. Wright and P. Ayton.

reported since Lichtenstein et al.'s review, excepting that which, in our view, provides either strong support for, or militates against, a particular model.

## 18.2 THE OVERCONFIDENCE EFFECT AND THE HARD–EASY EFFECT

Errors in probability judgements are not randomly distributed around the "target" value (e.g. the normative answer derived from the application of Bayes' theorem, or the proportion of correct answers associated with a particular subjective probability judgement). Rather, systematic errors or "biases" are frequently observed in a variety of tasks requiring individuals to produce probability judgements. So robust and compelling are these effects that they have been christened "cognitive illusions" by a number of authors (e.g. Kahneman & Tversky, 1982; von Winterfeldt & Edwards, 1986). In the calibration literature, the most commonly observed bias is the overconfidence effect. Subjects deliver probability estimates which are too high when measured against either the relative frequency of occurrence of an event assigned a particular probability estimate, or the proportion correct of answers which have been assigned a particular probability value.

A closely related phenomenon is the hard–easy effect; this is the observation that overconfidence decreases as task difficulty (usually indexed by the overall proportion correct) decreases. With easy tasks (over about 80% correct answers on a half-range probability scale, in a *two-alternative forced-choice* [2AFC] task[1]) the overconfidence effect disappears, and underconfidence is often observed. The main challenge facing theories and models of confidence is to explain these two effects.[2] It should be noted however, that there are clear anomalies in the literature. Good calibration has been found for "difficult" tasks (e.g. Keren, 1988); marked differences in calibration performance have been observed at the *same* level of task difficulty (e.g. McClelland, Coulson & Icke, 1990; Wright, 1982) as has good calibration at *different* levels of item difficulty (Juslin, 1993). Reversals of the hard–easy effect have also been noted, where the degree of overconfidence for a harder task is less than for an easier task (Keren, 1988; Ronis & Yates, 1987). Any principled account of how individuals make subjective probability judgements has to provide an explanation for these results, as well as the overconfidence effect and the hard–easy effect.

## 18.3 THE LOCUS OF BIAS IN PROBABILITY JUDGEMENTS

Over the last twenty years or so, two rival schools have developed, each with



a radically different view as to the locus of the observed biases in calibration and other probability tasks. Jungermann (1983) termed one camp "the pessimists" and the other "the optimists". The pessimists believe that biases are in people—the optimists believe that biases are in research. The most representative members of the pessimist school are Daniel Kahneman and Amos Tversky. In their "heuristics and biases program" (Gigerenzer, 1991, page 85) they have argued that the locus of the bias is within the cognitive system, and have provided many demonstrations of the apparent irrationality of individuals when engaged in probabilistic reasoning (e.g. Kahneman, Slovic & Tversky, 1982; Tversky & Kahneman, 1974, 1983). Kahneman and Tversky claim that an explanation for this irrationality is that individuals use a variety of heuristics when reasoning probabilistically. Several of these heuristics have been cited as explanations (or partial explanations) for miscalibration, most notably the "anchor-and-adjust" heuristic. For example, Keren (1991) proposed that in laboratory-based 2AFC tasks, the expectation of the subjects regarding task difficulty might act as an anchor, and explain the relationship between difficulty and over/underconfidence. He suggested that subjects might anchor on a probability estimate reflecting intermediate difficulty (75%). When confronted with an item perceived to be either very easy or very difficult, they would adjust accordingly, but not sufficiently, and this would lead to under- or overconfidence respectively. Ferrell & McGoey (1986) made a similar suggestion. Wright (1982) also appealed to the anchor-and-adjust heuristic in order to explain the difference in calibration for past-event questions (e.g. has at least one member of the British Parliament died within the last fourteen days? (a) yes, (b) no) and future-event questions (e.g. will at least one member of the British Parliament die within the next fourteen days? (a) yes, (b) no). He suggested that the response anchor for past event questions might be 1.0 (reflecting certainty) whereas for future event questions it might be 0.5 (reflecting uncertainty). A failure to adjust sufficiently from these anchors would lead to the observed overconfidence for past-event questions and underconfidence for future-event questions.

Other theorists have sought explanations for miscalibration (and particularly overconfidence) in terms of cognitive style (Wright & Phillips, 1984), ignorance of processing limitations (Pitz, 1974), motivation (Milburn, 1978; Zakay, 1983), cognitive optimism (Dawes, 1980), and response-scale effects (Poulton, 1989). However, as Keren (1991) noted, many of these explanations are *post hoc* in nature, and whilst most of them are consistent with the finding of overconfidence, they cannot explain observations of underconfidence, good calibration and the hard–easy effect. The feature they have in common is that they all attribute miscalibration to human failing.

The most vigorous champion of the optimist school is Gerd Gigerenzer. In a number of papers (Gigerenzer 1991, this volume; Gigerenzer, Hoffrage & Kleinbölting, 1991) he and his colleagues have argued strongly against the

pessimist school, and in particular the heuristics and biases approach. They have provided both a theoretical framework and empirical evidence to back the claim that biases in probabilistic reasoning are essentially artifacts, encouraged by the use of artificial and sometimes misleading tasks, and the nonrepresentative sampling of stimulus materials. In addition to suggesting that the locus of bias is in the main outside the cognitive system, Gigerenzer has also questioned the nature of probabilistic representations *within* the cognitive system. He has argued that the "intuitive statistician" within us is a frequentist, and not a Bayesian. Thus for Gigerenzer, probabilities are represented in terms of frequencies—and not as beliefs. This, Gigerenzer argues, has profound consequences both for the interpretation of the empirical evidence and for the nature of the theories and models required to explain human probability judgement, as we will discuss in a later section.

## 18.4 THEORIES AND MODELS

In this section, each of the theories and models we have chosen to review is briefly described, and in the next section critically evaluated. Although not exhaustive, we hope that we have included most of the major theoretical work from 1980 to date. Some of the models are clearly within the pessimist camp (locating the bias within the individual) and others the optimist camp (locating the bias within the experimental procedure, and in particular the nature of the stimulus materials). In addition, it is clear that some are domain specific (e.g. restricted to general-knowledge tasks), whereas others are presented as quite general models. Some of the models seem more applicable to situations in which the stimulus items are *essentially similar* (Ronis & Yates, 1987) or *related* (Keren, 1987, 1991) that is, they share common characteristics, whereas others are applicable to situations in which the stimulus items are *essentially unique* or *unrelated* (Keren, 1991). Despite these differences, we attempt a comparison of the models in a later section. The models are presented in chronological order.

### 18.4.1 The Stage Model

Koriat, Lichtenstein & Fischhoff (1980) proposed a three-stage model of the cognitive processes involved in answering a two-alternative general-knowledge question, and giving an associated confidence rating. In the first stage, memory is searched for relevant information and an answer chosen; in the second stage the evidence is assessed to arrive at a feeling of certainty, and in the third stage this feeling is translated into a numerical response. Koriat et al. suggested that unwarranted certainty (overconfidence) might be linked to one or more of the three stages.



They proposed that in the first stage, individuals might be biased in the way they elicit knowledge, favouring positive rather than negative evidence. In the second stage, they suggested that individuals might have a tendency to disregard evidence inconsistent with the chosen answer. This tendency to elicit positive evidence, and disregard evidence contrary to the chosen alternative, would lead to overconfidence. Finally, they noted that in addition to the cognitive biases operating at the first two stages, there might be an inappropriate translation of feelings of certainty into a probability value. If this mistranslation were such that individuals gave values that were generally too high, this would also contribute to the overconfidence effect.

### 18.4.2 The Detection Model

Ferrell & McGoey (1980) proposed a model for calibration based on signal detection theory (also see Ferrell, this volume, and Smith & Ferrell, 1983). These authors not only provided an explanation of how perceived truth of propositions might be translated into numerical judgements of confidence (corresponding to the third stage in Koriat, Lichtenstein and Fischhoff's model), but also sought to explain the overconfidence effect, the hard–easy effect, and the effects of base-rate change on calibration performance.

Ferrell and McGoey suggested that the task facing subjects in a calibration study can be broken down into two parts; the first being a detection process, described by a signal detection model, and the second the assignment of a numerical probability value on the basis of the result from the first stage. The decision variable used is partitioned by a set of criterion values, one interval for each possible probability response ($r$). Each question generates a particular value on the decision variable, and the interval into which that value falls then determines the numerical response. In a 2AFC task, it is assumed that each alternative produces a value of apparent truth, and the subject chooses the alternative with the higher value. For simplicity, the distributions of apparent truth for the two alternatives are assumed to be normally distributed with equal variance. The decision variable is then taken to be the *absolute difference* between the truth values for the two alternatives, the larger the difference the greater the confidence that the correct alternative has been selected. The distributions of absolute difference when the correct answer produced the larger truth value, and when the incorrect answer produced it, are normal distributions truncated below zero. Calibration can then be determined from (1) the probability of a correct response [p(C)], (2) the cumulative distribution functions of the decision variable when the response is correct and not correct, and (3) the partition of the decision variable.

In order to be perfectly calibrated, subjects must choose a partitioning such that for each interval, $p(C \mid r) = r$. However, Ferrell and McGoey assume that the partition is determined by information obtained prior to the task (subjects

appear to set their criteria for a task of intermediate difficulty, i.e. about 75% correct—see Ferrell & McGoey, 1980, page 40) and that it will *not change* unless feedback about performance is provided. In a two-alternative task, the partitioning should be determined solely by discriminability (as base rate is fixed at 50%). If subjects have set their partitioning for a task which is *easier* than the task actually presented, they will exhibit overconfidence—if on the other hand, the partitioning is set for a task *harder* than the one presented they will exhibit underconfidence. Ferrell and McGoey also showed that the model is not restricted to 2AFC tasks, but can be applied quite generally to any calibration task format. Finally, with the assumption that subjects are insensitive to changes in base rate as well as discriminability, they also provided predictions concerning the effects of base rate change on calibration performance in full-range tasks (see also Smith & Ferrell, 1983, and Ferrell, this volume for further details of the model).

### 18.4.3 The Process Model

May (1986a, 1986b) proposed a process model of subjective probability judgements which she claimed would allow the degree and direction of miscalibration to be predicted. In common with the later ecological models (see below) she argued that miscalibration should neither be seen as a bias in inferential reasoning, nor as a result of mistranslation of a feeling of uncertainty into a numerical response (cf. Ferrell & McGoey, 1980). Instead, she suggested that it was a consequence of the specific background knowledge possessed by subjects, of the tasks given to subjects, and the selection of items within the tasks.

Following Koriat, Lichtenstein & Fischhoff (1980) her model has three stages, which she labelled *problem-solving*, *emergence of subjective certainty* and *quantification* respectively, but she placed the origin of miscalibration at the first stage. She also identified two sources of difficulty which would affect the proportion of correct responses in a calibration task. The first source (Difficulty 1) was seen as a characteristic of the task such as the objective distance between two stimuli in a psychophysical task. The second source (Difficulty 2) was attributed to the subject having "wrong knowledge" (such as a distorted cognitive map when making a latitude judgement).

In May (1986a) two possible internal representations based on a subject's knowledge are presented. The first *mental model* is in the form of a syllogism, which May proposed might be used to answer a question such as "Which city has more inhabitants? (a) Hyderabad, (b) Islamabad." The second is in the form of a cognitive map, and she suggested that this type of representation might be used to answer a question such as "Which city is further north? (a) Rome, (b) New York."



With the syllogistic representation, May speculated that a subject might reason in the following way; Capital cities tend to have many inhabitants, Islamabad is a capital city, and therefore Islamabad is likely to have many inhabitants. On this basis, the subject would choose Islamabad as the answer.[3] May proposed that the confidence expressed by the subject would be a function of the perceived extent of the intersection between the set of capital cities and the set of cities with a large population, and possibly other relevant background knowledge. She argued that if the items (pairs of cities) were *randomly* sampled from the population of cities, good calibration would be expected in the long run, but if a set of items were selected so as to include a large number of "misleading" items (i.e. items for which the inference produced the wrong answer—as in the example above), overconfidence would result.

With the cognitive-map representation, May proposed that the difficulty of an item (e.g. deciding which of two cities was further north) would depend upon the *subjective* distance between them, and this would be reflected directly in the confidence given. However, distortions in subjects' cognitive maps could lead them to pick the wrong alternative, and depending on the extent of the distortion, to pick the wrong alternative with considerable confidence. Again, a large number of such "misleading" items in a set would produce overconfidence. In the cognitive-map representation, the probability of a correct answer (the *reaction* probability in May's terminology) is determined by the *subjective relationship* between the cities (i.e. which is subjectively further north) and the confidence (or subjective probability) by the *subjective distance* between the cities.

Finally, May draws a distinction between *populationwise* calibration and *itemwise* calibration. The former defines calibration for the universe of items that could be constructed within a certain knowledge domain. She suggests that defined this way, perfect calibration is impossible when misleading items are present. Itemwise calibration is defined with respect to single items, so that an item is calibrated when the mean reaction probability is identical to the mean subjective probability. Thus, by definition, only non-misleading items could be well calibrated in this sense.

### 18.4.4 The Memory Trace Model

Albert & Sponsler (1989) presented a mathematical model for the calibration of subjective probabilities. They proposed that the brain subconsciously makes subjective probability estimates based upon memories of similar prior experiences.

The basic principles underlying the model are fairly simple. Albert & Sponsler (1989) assume that when confronted with a new "fact pattern", the brain abstracts cues which permit it to identify a set of prior experiences

characterized by a similar set of cues. For example, a weather forecaster might, on the basis of current weather conditions, identify days in the past upon which a similar pattern of weather conditions prevailed. The memory trace which is retrieved is considered to be composed of the results of *predictions* of a series of binary events (e.g. rain, no rain) which either did or did not occur. Successful predictions are imagined to be encoded as 1s and unsuccessful predictions as 0s. The "true" subjective probability is the relative frequency of successful predictions in the past for the entire set of events.

Albert and Sponsler suggested that an individual may not be able to retrieve the entire set of previous predictions, but rather a subset of the entire memory trace. The particular subjective probability estimate is then taken to be the relative frequency of successful predictions within the subset identified. An expert[4] estimator, according to the authors, will be able to identify the full set, and thus his or her estimate will match the true subjective probability. A less expert estimator will be able to retrieve only a subset of the full set of prior predictions, and thus his or her prediction will not necessarily correspond to the true subjective probability (the actual estimate depending upon the relative frequency of successes in the subset).

From this basic model, the authors derive the permissible range of subjective probability estimates allowed by the model for various values of the true subjective probability, and for various proportions of the memory trace retrieved. The range of possible subjective probability estimates a subject *could* produce is constrained (according to the model) by both the proportion of successful predictions in the full memory trace, and by the proportion of the trace retrieved on a given occasion. The authors assumed that, on average, the subjective probability estimate given is the midpoint of the range of permitted values for a given value of the true subjective probability and a given proportion retrieved. The rationale for this hypothesis is simply that when the midpoints are plotted against the true subjective probability values for various proportions retrieved, *overconfidence* is observed in that the midpoint values are greater than the corresponding true probabilities. The authors also conclude that the choice of subsets cannot be random (such that all subsets of a particular size are equally likely to be retrieved) as they show that this leads to an expected value of the subjective probability estimates which is equal to the true subjective probability value for all sample sizes. In other words, if the choice was random, subjects would in the long run be perfectly calibrated, and not demonstrate overconfidence.

In the remainder of their paper the authors speculated as to the possible shape of the distribution of the probability estimates in the various permitted ranges, and suggested that a transformation of the Beta distribution was particularly promising. They discussed the problems with attempting to estimate empirically the parameters in their model, and this leads to further speculation concerning the possible shape of the distribution of estimated true



subjective probabilities. Two equations are given to calculate the expected value of the true probability estimates, one assuming a uniform distribution of values over the permitted range, and the other for a non-uniform distribution. The authors also note that an estimate of the true subjective probability could be obtained directly from an individual's calibration curve. Albert & Sponsler (1989) conclude that "... the entire theory demands, and it is hoped will receive, experimental verification" (page 308).

### 18.4.5 The Ecological Models

Working independently, Gigerenzer, Hoffrage & Kleinbölting (1991), and Juslin (1993, 1994) produced two models of remarkable similarity. We have termed these the "ecological" models. Both of these models are founded on the simple but powerful notion that, as a result of interaction with the natural environment, individuals encode the frequencies of co-occurrences of events in the environment, and use this information in a very direct fashion when making judgements about discrete propositions and attaching confidence ratings to those judgements.

The theory of probabilistic mental models (PMM theory) proposed by Gigerenzer, Hoffrage & Kleinbölting (1991) was developed to explain performance in 2AFC general-knowledge tasks, but the authors do suggest that the model could also be applied to 2AFC perceptual tasks. The authors outline the circumstances under which good calibration is to be expected, and provide explanations for the overconfidence effect, the hard–easy effect, observed reversals of the hard–easy effect, and a hitherto unobserved third phenomenon, termed the confidence–frequency effect (described below).

Underlying the model are the following assumptions;

(1) individuals are well adapted to their environments (see Brunswick, 1943, 1955),
(2) individuals are able to extract and store accurately, information regarding the frequency of occurrence of events in the environment—and do so with little if any conscious effort (see Hasher and Zacks, 1984 for a review),
(3) the basis for probability judgements are these stored frequencies—the "intuitive statistician" is a frequentist.

Gigerenzer, Hoffrage & Kleinbölting (1991) propose that if a solution to a given general-knowledge item cannot be obtained directly (e.g. via direct retrieval from memory, or by use of an elementary logical operation such as the method of exclusion) the subject will set up a *probabilistic mental model* (PMM). To take an example used by Gigerenzer *et al.*, imagine that the task consists of deciding which of two German cities with more than 100 000 inhabitants (*a* or *b*) is the larger. The PMM will contain the reference class (all cities in Germany with more than 100 000 inhabitants), a target variable

(city size), probability cues (other variables related to the reference class) and cue validities. Gigerenzer et al. suggested that potential probability cues might include a soccer team cue (one city has a team in the Bundesliga and the other does not), an industrial cue (one city is located in an industrial area and the other a rural area), and a state capital cue (one city is a state capital and the other is not). A variable is a probability cue for the target variable in the reference class if the conditional probability of alternative *a* being the correct answer is different from the conditional probability of *b* being correct. For this example, a subject might use a soccer-team cue; that is choose the city which has a team in the German soccer Bundesliga (note that this assumes the cue can be activated; if both or neither of the cities had a team in the Bundesliga it could not be used). The *ecological validity* of this cue is 0.91, in that if all pairs in which one city has a team in the Bundesliga and the other does not are checked, one would find that in 91% of cases the city with the team in the Bundesliga has more inhabitants.

Cues are assumed to be generated, and if possible activated, in a hierarchical fashion. The probability cue with the highest validity is generated and tested first; if it can be activated it is used, if not, a further cue is generated and tested. If no cue can be activated, it is assumed that the subject chooses randomly and gives a 50% confidence rating.

Gigerenzer, Hoffrage and Kleinbölting argued that through interaction with the natural environment the ecological validities of cues become internalized through a process of observing the frequencies of co-occurrences of environmental events, and become the *cue validities* in the PMM. An individual uses a probability cue to both select an answer, and as the source of the confidence; once a choice is made, the cue validity is given as the confidence rating. If individuals have had repeated experience with a particular reference class, a target variable, and cues in the environment, it is assumed that the cue validities correspond well to the ecological validities. However, if a subject in a calibration experiment is given a set of items which are not *representative* of the reference class in the environment, performance will be systematically biased, as the cue validities used will not be appropriate.

Gigerenzer, Hoffrage & Kleinbölting (1991) made a number of predictions based on PMM theory. The first was that typical general-knowledge items (which have been used extensively in calibration studies) will produce both overconfidence and accurate judgements concerning the number of items correctly answered (frequency judgements). Overconfidence is attributed to a biased selection of items, with difficult, and importantly, "misleading" items, being over-represented. The use of cues and cue validities which would produce good calibration for a representative set of items leads to overconfidence with a selected "difficult" set. Imagine that a subject uses a cue with an ecological validity of 0.90 to answer ten questions. If these have been sampled randomly, the set of ten questions would be expected to contain one misleading item—an item where the cue fails to deliver the correct answer. The subject would be expected to get 9 correct, and give a confidence of 0.90 for each answer. If however the set has been informally selected, there may be, say, 3 such misleading items present, in which case the subject would only get 7 correct, but would still give a 0.90 confidence rating for each of the 10 questions. Hence the subject would be overconfident. However, the authors also predicted that if subjects were asked "how many items do you think you answered correctly?" they should give accurate estimates, as the reference class is now past general-knowledge tests that they have taken. If the current test is typical of the general-knowledge tests they have experienced in the past (i.e. representative of the reference class of general-knowledge tests), good calibration with respect to frequency estimates is to be expected. A corollary to the first prediction is that if subjects are given a set of items randomly selected from a particular reference class, they will exhibit good calibration with respect to confidence judgements, but should now *underestimate* the number of items correctly answered. Gigerenzer, Hoffrage & Kleinbölting (1991) referred to this as the *confidence–frequency effect*. They further predict that if two sets of items, hard and easy, are generated by the *same* sampling process (be it random or biased) the hard–easy effect should disappear. Finally, they predicted that if a set of items is representative of a "hard" reference class, and a second set is selected to be "difficult" but from an "easy" reference class, a reversal of the hard–easy effect should be observed. Empirical evidence is presented (and results in the literature reinterpreted) which, in the main, support the predictions from the model.



The arguments developed by Juslin (1993, 1994) to explain the overconfidence effect and the hard–easy effect are, in all essentials, the same as those of Gigerenzer, Hoffrage & Kleinbölting (1991). However, Juslin did not make the further predictions concerning the calibration of frequency judgements, nor did he predict the possible reversal of the hard–easy effect.

### 18.4.6 The Strength and Weight Model

Griffin & Tversky (1992) have provided the most complete model within the heuristics and biases program to explain the patterns of overconfidence and underconfidence observed not only in calibration studies, but also in other investigations of judgement under uncertainty. In this respect, it has similarities with Gigerenzer's (1991) attempt to explain apparent biases in many situations using a small number of explanatory principles.

The two concepts central to Griffin & Tversky's (1992) argument are those of "strength" and "weight". Neither of these concepts is rigidly defined, but by strength they mean the "extremeness" of available evidence, and by weight the "predictive validity" of the evidence. They note that the distinction between these two concepts is closely related to the distinction between the size

of a statistical effect (e.g. the difference between two means) and its reliability (e.g. the standard error of the difference). Griffin and Tversky argue that individuals focus on the strength of evidence and may make some adjustment (albeit insufficient) in response to the weight. This thesis makes particular use of two of the heuristics identified in the heuristics and biases program "representativeness" and "anchor-and-adjust". For example, individuals may make use of the representativeness heuristic (judging an interviewee on how much he or she "looks like" a successful manager) whilst ignoring (or paying scant attention to) other factors controlling predictive validity. Any adjustment that is made to take into account the weight of evidence is deemed insufficient; i.e. individuals make use of the anchor-and-adjust heuristic, but fail to adjust sufficiently.

Griffin and Tversky claimed that their hypothesis predicts a distinctive pattern of overconfidence and underconfidence. When, in a given situation, strength is high but weight is low, subjects exhibit *overconfidence*. However, when strength is low and weight is high, individuals should exhibit *underconfidence*. In the first half of the paper, they were concerned with testing their predictions with respect to the evaluation of statistical hypotheses; in the second half of the paper they extended their argument to confidence judgements, and in particular to the calibration of general-knowledge questions. The authors noted that there is a problem with the application of the theory within the calibration domain, as strength and weight cannot be experimentally controlled. However, Griffin and Tversky offer an "analogy to a chance setup" (page 425) as a model of the processes involved when making confidence judgements in a calibration study.

In this model, the balance of arguments for a (two-alternative) general-knowledge problem is represented by the proportion of red and white balls in a sample; difficulty (discriminability) is the difference between the probability of obtaining a red ball under each of two competing hypotheses (the correct alternative and the incorrect alternative). Expressed confidence is given by the balance of arguments, i.e. the proportion of red balls in the sample (where a red ball represents an argument in favour of the correct alternative). For any given sample size, and any pair of probabilities of obtaining a red ball under the competing hypotheses, the normative or "correct" confidence response can be computed for each sample composition (i.e. 1 red, 2 reds, etc.) from the Binomial distribution and the application of Bayes' theorem. Griffin and Tversky assume that the confidence judgement given by an individual is simply the proportion of red balls they observe (i.e. the strength of the evidence). Thus neither the level of difficulty as indexed by the discriminability of the hypotheses, nor the sample size (both aspects of the weight of the evidence) are taken into account. In a simulation of the model, the authors generated three calibration curves, by plotting the normative (posterior) probability solution against the proportion of red balls in the sample (balance of



arguments) for three pairs of hypotheses defining three levels of difficulty. The probabilities of obtaining a red ball under the competing hypotheses were 0.50 and 0.40 for the "easy" task, 0.50 and 0.45 for the "difficult" task, and 0.50 under both hypotheses for the "impossible" task. Griffin and Tversky chose non-symmetrical hypotheses "to allow for an initial bias that is often observed in calibration data" (page 426).[5] The sample size (10) was held constant. The three curves bore a striking resemblance to empirical calibration curves obtained by Lichtenstein & Fischhoff (1977) for three levels of item difficulty (as indexed by overall proportion correct). Calibration was reasonably good for easy items (with slight underconfidence for lower confidence ratings and slight overconfidence for higher ratings), there was marked overconfidence for difficult items, and a flat calibration curve for impossible items. Thus the model appears to provide an explanation for both the overconfidence phenomenon and the hard–easy effect. It would also be an easy matter to simulate changes in base rate with this model, but this was not investigated by the authors.

## 18.5 EVALUATION OF THE MODELS

In this section, we provide a critical evaluation of the models described above. We assess each model with respect to the empirical evidence, and in terms of psychological plausibility. We also try to highlight the similarities and differences between the models.

### 18.5.1 The Stage Model

Koriat, Lichtenstein & Fischhoff (1980) presented some empirical evidence to support the view that overconfidence can be attributed to biases operating at the first and second stages of their model. They noted a bias in the production of reasons for and against a particular alternative, favouring evidence *for* over evidence *against*. They also provided some empirical support for the notion that subjects disregard evidence inconsistent with their chosen answer. Forcing subjects to write down a contradictory reason did improve the realism of their confidence assessments as indexed by calibration scores. However, the decrease in overconfidence was very small (2%) and non-significant. In a replication of the study Fischhoff & McGregor (1982) failed to find an effect of disconfirming evidence. Gigerenzer, Hoffrage & Kleinbölting (1991, page 521) argued that these negative results were consistent with PMM theory, which predicts no change in expressed confidence when subjects are asked to produce disconfirming evidence.

Despite the underspecification of the stage model, and the lack of empirical support in its favour, it is useful when viewed as a framework within which

other models of calibration can be located. For example, the process and ecological models are concerned with the first stage, the strength and weight model attributes miscalibration to the first and second stages, and the detection model to the third stage.

### 18.5.2 The Detection Model

Ferrell and his colleagues (Ferrell, this volume; Ferrell & McGoey, 1980; Smith & Ferrell, 1983) have provided an impressive amount of evidence in support of their model. They have shown that it provides a good fit to a wide range of data sets, collected using a variety of task formats and types of stimulus materials.

With respect to task difficulty in the standard 2AFC probability-correct task, Ferrell & McGoey (1980) have shown that even with the estimation of a single parameter [$p(C)$], both the calibration curve and the usage of the response categories can be well predicted. For example, they estimated the cutoffs (the position of the criteria on the decision variable) for the entire data set collected by Lichtenstein & Fischhoff (1977), and showed that this set of criteria provided a good fit to the data for subsets of items (e.g. "hard" items and "easy" items). They concluded that such findings were consistent with the hypothesis that a set of cutoffs appropriate to a proportion correct of about 75% was maintained even when $p(C)$ was substantially different from 75%. The failure to adjust the cutoffs (or to adjust them sufficiently) led to the hard–easy effect.

Ferrell & McGoey (1980) and Smith & Ferrell (1983) also showed that in a full-range probability-true task, the shift of the calibration curve under different base-rate conditions was again predictable if the cutoffs for the 50% condition were used to model the data collected with base rates either above or below 50%. Again, they concluded that the effect could be attributed to the subjects' failure to adjust their criteria with changes in base rate.

Despite the success of the detection model,[6] it has been criticized on the grounds that it does not elucidate the cognitive processes involved in making subjective probability judgements. For example, Keren (1991, page 262) remarked that "Unfortunately, the model provides little insight into the possible cognitive processes governing probability judgements." Of course, this does not mean that the model is wrong—it could indeed be the case that miscalibration is caused by a problem with translating a feeling of certainty into a numerical estimate, and has little to do with the use of heuristics, or the operation of other cognitive processes.

However, there are some empirical results which suggest that calibration performance may depend on more than the numerical assessment process. For example, Juslin (1993) found excellent calibration in four subsets of data where $p(C)$ varied from 66% to 80%. Thus the subjects in this experiment



were able to maintain good calibration at rather different levels of difficulty—a finding inconsistent with the predictions of the detection model. Further, a fundamental assumption underlying the predictions derived from the model is that when miscalibration is observed, subjects have set their criteria for a task which they believe to be either easier (leading to overconfidence) or harder (leading to underconfidence) than the task actually presented. However, Gigerenzer, Hoffrage & Kleinbölting (1991) have presented convincing evidence that subjects are apparently able to anticipate very accurately the difficulty of a *typical* general-knowledge test (i.e. one representative of the reference class "general-knowledge tests") in that they can give good estimates of the numbers of items they have correctly answered, but still remain very overconfident in their calibration of individual items. In addition, with a randomly sampled set of items, subjects underestimate the frequency of correct answers but show good calibration. Both of these findings present problems for the detection model; with task difficulty correctly determined good calibration would be predicted, and when task difficulty is overestimated, underconfidence should be the result.

In conclusion, the detection model can be regarded as being a model of the last of Koriat, Lichtenstein & Fischhoff's (1980) three stages—the stage where subjective feelings of uncertainty are mapped onto numeric probability responses. The detection model has little to say about the cognitive processes leading to the formation of these feelings other than that some feature (or features) of the task generates (in an unspecified manner) a value on an unscaled internal variable (the decision variable). By making the simple assumptions that probability responses are read off from a partitioning of this variable and that—*in the absence of feedback*—this partitioning is not appropriately matched to the task difficulty the detection model can account for empirical data from a number of domains. This fact suggests that in most calibration tasks we need not look at earlier stages in Koriat et al.'s framework in order to account for the observed phenomena. However, for other tasks we may need to look further. For example, as emphasized above, the detection model makes no prediction of miscalibration when feedback is present. In the majority of judgement tasks outside the laboratory feedback of some kind is available, but in a number of such tasks miscalibration has still been found (e.g. Staël von Holstein, 1971, 1972; Yates, 1982; Yates and Curley, 1985). Further, the detection model is not applicable to tasks where explicit reasoning about numbers or proportions is required, such as in Bayesian probability revision or "book-bag-and-poker-chip" experiments. In these tasks—which Ferrell refers to as "external validity tasks" (see Chapter 17)—no value is generated on the internal decision variable, hence no probability response can be generated as required by the detection model. It would seem then that the detection model can only provide a partial explanation for poor calibration performance.

### 18.5.3 The Process Model

The two papers outlining the process model (May, 1986a, 1986b) contain many interesting and novel ideas, some of which were taken up in the later ecological models. For example, May recognized that the degree of miscalibration observed in a particular general-knowledge task was likely to depend upon the nature of the selection process used to generate the items, and she argued against the notion that miscalibration could be attributed to shortcomings in human inferential or intellectual reasoning. Finally, she introduced the notion of the *mental model* as the basis for subjective probability judgement.

May proposed that there might be at least two types of representation that could be used to answer general-knowledge problems. With respect to the syllogistic mental model where the individual uses inference to choose between two alternatives in a 2AFC item, Gigerenzer, Hoffrage & Kleinbölting (1991) have argued that the probabilistic syllogism as presented by May would not lead to good calibration in the long run, because it did not include information about both the alternative answers. They proposed a modified version of the model (the *double-syllogism model*) which, they claimed, would result in long-run calibration (see Gigerenzer, Hoffrage & Kleinbölting, 1991, page 523).

For the second form of representation (a cognitive map) May argued that perfect calibration was impossible if the map contained distortions. With this form of representation, misleading items are misleading because of "false knowledge" possessed by subjects. She showed that confidence was highly correlated not with the *objective* distances and geographical relationships between cities, but with the *subjective* distances and relationships, and that these were distorted. Thus she argued that the reason that subjects gave a mean confidence rating of 80% when answering the question "Which is further North? (a) Rome, (b) New York", though the solution probability was somewhat under 30%, was because of a serious distortion in the subjects' cognitive map, with North American cities shifted too far North with respect to European cities.

However, a simpler explanation can be derived from the ecological models—subjects used a climate cue to answer the question knowing that in general, a colder climate indicates a higher latitude. This would lead them to pick the wrong answer (i.e. New York) and may also lead them to believe that New York *really* is further North than Rome—resulting in a distorted cognitive map. The subjects had obviously never seen a map on which the latitudes of these two cities are reversed, so why did they have distorted cognitive maps? The idea that subjects use probability cues in answering such questions supplies an answer to both why subjects get this item wrong with high confidence, and why they have distorted cognitive maps. Differences in the confidence expressed for different pairs of items can be attributed to the use of a variety of cues with varying cue probabilities.



Although for the example above, the use of probability cues may provide a more parsimonious explanation of calibration performance, May is not alone in believing that different representations might be used in different stimulus domains. Björkman, Juslin and Winman (1993) argue that for psychophysical judgements, the distance between stimulus items on the dimension to be judged is indeed the representation used—but believe that for general-knowledge items probability cues are used.

May was clearly wrong in believing that perfect calibration was only possible in the absence of "misleading" items (see Gigerenzer, Hoffrage & Kleinbölting 1991; Juslin, 1993, 1994), and for the general-knowledge tasks that she considered, the ecological models would seem to provide a better description of the cognitive processes and representations underlying subjective probability judgements than is furnished by her process model.

### 18.5.3 The Memory Trace Model

There are a number of problems with this model as a general explanation of calibration performance. Firstly, it would only seem applicable to situations in which the assessor has had past experience of a set of similar events to those presented at test, *and* has received outcome feedback (e.g. a weather forecaster predicting rain). The model does not seem applicable to tests where the items are *essentially unique* (Keren, 1991) or indeed any task which is novel (e.g. choosing which of a pair of countries has the larger population) although the stimulus domain may be familiar (e.g. countries of the world). Indeed, Albert and Sponsler (1989) stated that confidence is based on a record of past successes and failures at predicting the outcomes of events the brain "deems similar" (page 298).

Secondly, the model only predicts the overconfidence effect—indeed, the authors seem to have been unaware of the fact that with very easy tests underconfidence is observed. Their apparent belief that it is only overconfidence which has to be explained (and their faith in the robustness of this finding) was critical to their rejecting the notion that all subsets of the full memory trace are equally likely to be selected because this would, in the long run, lead to perfect calibration. According to this model, perfect calibration can only be achieved when the assessor retrieves the entire memory trace—implying that only an assessor with perfect memory for the outcomes of the predictions can be perfectly calibrated.

The model does, however, have some similarities with the ecological models, in that the cognitive representation which guides both the decision and subjective probability estimate is in terms of frequencies. However, unlike the ecological models, the frequencies simply represent past successes and failures at prediction, and not the validities of various probability cues associated with a particular target variable and a particular reference class.

Finally, no distinction is drawn between performance in terms of relative frequency of success in the past, and confidence in individual items. Gigerenzer, Hoffrage and Kleinbölting (1991) have drawn a distinction between the reference class of past success on similar tests, and the reference class relating to the content of individual items, and have provided evidence for the psychological reality of this distinction. In the memory trace model they are one and the same. If subjects can produce good estimates of the frequency of success in the past they should also be well calibrated—but Gigerenzer et al. have shown that this is not so.

In summary, this is a model designed to explain individual differences in calibration performance (cf. Phillips & Wright, 1977) but suggests that differences in the calibration performance of difference assessors is entirely attributable to the quality of their memories. It fails to capture many of the empirical findings relating to calibration performance, and would seem to be an implausible candidate for either a general or domain-specific explanation of subjective probability judgement.

### 18.5.4 The Ecological Models

The PMM theory described by Gigerenzer, Hoffrage and Kleinbölting (1991) is the most complete model for the calibration of subjective probabilities that has so far been produced. It elegantly explains the overconfidence effect, the hard–easy effect, the circumstances under which good calibration is to be expected, and the confidence–frequency effect. It also makes strong and testable predictions concerning the circumstances under which a reversal of the hard–easy effect should occur. The model also explains a number of other apparently anomalous findings in the literature. The authors argue that the locus of miscalibration for general-knowledge items is in the test materials themselves, and is not the result of biased probabilistic reasoning on the part of the subjects.

However, the empirical evidence presented by Gigerenzer, Hoffrage & Kleinbölting (1991) is somewhat less convincing than the model. The problem is that difficulty as indexed by proportion correct co-varies with the type of item selection—items which are selected to be a good test of an individual's general knowledge (and thus not representative of the reference class) will on average be harder than those randomly selected from the reference class. Thus the demonstration that calibration for randomly selected city items is better than for standard general-knowledge items could be viewed as just another example of the hard–easy effect. However, the finding that subjects can be overconfident with respect to calibration based on the confidence expressed for individual items, and simultaneously well calibrated with respect to their overall performance with informally selected items, or well calibrated for individual ratings and underconfident about their overall performance with randomly selected items, is much more compelling.



The model proposed by Juslin (1993, 1994) based on *internal cue theory* (Björkman, in press), is in most respects identical to the PMM theory. Juslin's model was restricted to an explanation of the overconfidence effect and the hard–easy effect, and he did not predict the confidence–frequency effect. However, Juslin (1993) provided an impressive empirical test of his own model, and thereby, PMM theory. Juslin predicted that if the *randomly* generated geography items used in his experiment were divided into four subgroups, not on the basis of proportion correct (as in Lichtenstein & Fischhoff, 1977) but on the basis of the *mean familiarity rating* given to the pair of countries forming an item, the hard–easy effect would be abolished, and good calibration should be observed—despite differences in the proportion of items correct across the subgroups. The reasoning was as follows: for highly familiar items, a large number of relevant cues can be generated, and thus there is a high probability of a cue with a high validity being activated. This will lead to a high proportion of correct answers. For items with low familiarity the reverse is true; they are likely to be answered using cues with low validities, and thus a low proportion correct is to be expected. In both cases, the cues used should be ecologically valid, as the generating process was random and thus good calibration would be expected. This prediction was supported by the data; for the most familiar items the proportion correct was 0.80 and for the least familiar 0.66, but for all subgroups calibration was excellent. Hence Juslin (1993) successfully decoupled cue appropriateness and item difficulty.

We have identified two potential problems with the ecological models. The first concerns the degree to which the models can be extended beyond the domain of knowledge questions, and beyond the 2AFC task format. Gigerenzer, et al. argued that PMM theory was applicable to perceptual tasks, and that good calibration would be anticipated as long as the items were not chosen to be misleading (i.e. not selected for perceptual illusions). They also predicted that with two perceptual tasks, which varied in discriminability but with stimuli generated by the same sampling process, the hard–easy effect should disappear. However, we have shown that overconfidence in a perceptual task varies systematically with discriminability (the hard–easy effect) despite the fact that the stimuli were indeed generated by the same *random* process (McClelland, Bolger & Tonks, 1992). It should be noted, however, that we used a full-range probability true task, and that the task was novel to the subjects, but nevertheless this finding does not square with either of Gigerenzer et al.'s predictions. With respect to the question of the generality of the probability cue notion, Juslin has taken an alternative approach, arguing that internal cue theory is only applicable within the knowledge domain, and that a different representation is used with psychophysical tasks (see Björkman, Juslin & Winman, 1993).

The second problem concerns the plausibility of individuals actually learning the appropriate cue validities for the probability cues with respect to

a target variable in a particular knowledge domain. To take Gigerenzer et al.'s example (although the argument also applies to Juslin, 1993, 1994) the frequency which the individuals would have to record is *the number of times that one city with more than 100 000 inhabitants has a larger population than another city with more that 100 000 inhabitants when the first city has a team in the Bundesliga and the second city does not*. Further to obtain an accurate cue validity, all possible pairs of cities would have to be examined, or at least to obtain an unbiased estimate, a random sample of all possible pairings would have to be selected. The appropriate cue validities could not be learnt if individuals merely noted that large cities *tend* to have teams in the Bundesliga, and smaller cities do not. Note also that if the target variable and probability cue are reversed (e.g. a decision has to be made as to which of two cities has a team in the Bundesliga, with population used as a cue) a different value would have to be recorded, as conditional probabilities are only symmetrical under very restricted circumstances. How plausible this is remains an open question.

In addition, both Harvey and Rawles (1992) and Griffin and Tversky (1992) have provided evidence inconsistent with the ecological models. Harvey and Rawles questioned the PMM assumption that subjects always choose the alternative with the higher value on the probability cue, and suggested instead that subjects "probability match" (Estes, 1964). Thus for a cue with a validity of 0.90, subjects would choose the alternative with the 0.90 probability 90% of the time, and the other alternative 10% of the time. Using a simulation technique, they found that probability matching model produced a very good fit to the data (from a general-knowledge test), whereas the PMM model gave a very poor fit.

The results inconsistent with the ecological models provided by Griffin and Tversky are described below.

### 18.5.5 The Strength and Weight Model

Griffin and Tversky (1992) presented both a general framework for understanding the relationship between confidence and accuracy, and a specific model for laboratory-based calibration experiments.

As described earlier, the specific model was in the form of an analogy (a chance setup) and the authors demonstrated that a plot of the "normative" solutions derived from Bayes' theorem against a measure of the strength of evidence (presumed to be the subjective probability estimates) produced calibration curves which mimicked the empirical curves from Lichtenstein and Fischhoff (1977).

It seems to us that this model is essentially a version of the detection model (Ferrell & McGoey, 1980) which makes use of a discrete probability distribution (the binomial distribution) rather than a continuous distribution (the



normal distribution). This view is shared by Ferrell (personal communication) who has developed another discrete version of the detection model (based on the symmetrical criterion model presented in Smith & Ferrell, 1983, pages 475–6). In this version, red balls represent evidence in favour of one of the alternatives in the 2AFC task (which may or may not be the correct answer) and white balls represent evidence in favour of the other alternative. If more red balls are present in the sample, the hypothetical subject would choose one alternative—if more white balls, the other. A sample containing exactly five red balls (out of 10) would lead to a random selection of an alternative, and a probability judgement of 0.50. This model produces slightly different curves from the Griffin and Tversky model, and the two models only coincide when the hypotheses are symmetrical. It should also be noted that the Griffin and Tversky model actually produces posterior probability values across the full probability range (from 0% to 100%) despite the fact that it is designed to be an analogy to a 2AFC task. To be consistent with Lichtenstein and Fischhoff's (1977) data, the authors are forced to "cut off" the calibration curves, and only plot values from 50% to 100%. The Ferrell version has the advantage that it does not predict confidence ratings below 50%, and thus the values fall in the half-range—as they should.

What are substantive differences between the strength and weight model and the detection model? Griffin and Tversky suggest that the strength of evidence is better represented by a balance of arguments, whereas Ferrell and McGoey suggest it is better represented by the absolute difference in apparent truth between the alternatives, measured on a continuous scale. The strength and weight model *never* allows for perfect calibration (even with feedback) as the proportion of balls in the sample is never the same as the posterior probability (except trivially, at the 50% point for an impossible task). The response criteria in this model are by necessity fixed (the number of red balls in the sample), whereas in the detection model it is possible for the criteria to be adjusted (with feedback) in order to improve calibration performance. In other respects, the models are very similar, in that they both assume that subjects base their probability judgements on the strength of evidence, and ignore the weight of evidence. Finally, if the sample size in the strength and weight model were allowed to tend to infinity, the binomial distributions would tend to normality and the probability scale would become continuous—as in the detection model.

In addition to the simulation of their model, Griffin and Tversky also reported some empirical results from an experiment (Griffin & Tversky, 1992, Study 5) which they interpreted as supporting the strength and weight approach, and as being inconsistent with PMM theory (Gigerenzer, Hoffrage & Kleinbölting, 1991). They showed that for a representative (*random*) sample of 30 pairs of American states, subjects were consistently *overconfident* in their predictions concerning population, high-school graduation rates, and the

difference in voting rates between the last two presidential elections. In addition, they had predicted that for population judgements both accuracy and confidence would be high (on the grounds that individuals should be knowledgeable about population) for voting both confidence and accuracy would be low (because they would not be knowledgeable about voting rates) and for education, accuracy would be low and confidence high. This last prediction was made on the grounds that subjects would be likely to use cues such as the number of famous universities or cultural events within a state to guide their judgements, when in reality the correlations between these cues and high-school graduations rates are very low—a type of "false knowledge" in May's terms (May, 1986a, 1986b). The predictions received empirical support. In particular, performance for both voting and high-school graduation was at chance level, although the mean confidence rating for education (65.6%) was significantly higher than that for voting (59.7%). The subjects were also asked to estimate how many of the questions they thought they had answered correctly, and it was found that for all three types of judgement the judged frequency was below the actual frequency (for voting and education the estimates were well below chance).

Griffin and Tversky concluded that overconfidence in calibration studies cannot be attributed to either an artifact of item selection or a by-product of task difficulty, and clearly their empirical findings would seem to pose a problem for the ecological models. However, Juslin (1993, 1994) had observed excellent calibration for population judgements—so the Griffin and Tversky result (6.5% overconfidence) for this attribute is somewhat of an anomaly. With respect to the other two attributes (high-school graduation and voting) performance was at chance level. This implies that these attributes were not part of the subjects' knowledge base (so effectively the task was impossible) but does not explain why subjects were overconfident (as the stimuli were *randomly* selected) or the difference in overconfidence between voting and education. However, the samples presented to subjects were very small (15 items per attribute) and could have contained a number of misleading items just by chance. Further, for solution probabilities around 50%, overconfidence would be expected simply because of the range restriction at the lower end of the probability scale (May, 1986b; Poulton, 1989). Finally, subjects are rarely faced with an impossible task, and may have suffered from a degree of *evaluation apprehension*, as well as wishing to be "good subjects" (McBurney, 1990). This may have led them to give confidence ratings higher than they truly felt appropriate, in an attempt to demonstrate that they could do the task. If the subjects were students (the source of the subjects is not stated) they might have felt that the experimenters would expect them to have knowledge concerning the education attribute in particular. This would lead them to provide higher confidence ratings for the education attribute than the voting attribute.



Griffin and Tversky's predictions do pay lip service to the notion that subjects use probability cues, but the implication is that subjects have "false knowledge"—they believe the cues to have higher validities than they actually do. Why this should be so is unclear from Griffin and Tversky's account.

### 18.5.6 Evaluation Summary

We have briefly described and critically reviewed seven models of subjective probability calibration. In the light of this review, should we be pessimistic or optimistic about the ability of individuals to be well calibrated?

Three of the models are clearly pessimistic; Albert and Sponsler (1989) suggest that overconfidence is a direct consequence of how the brain stores and retrieves information concerning past efforts at prediction. Both Koriat, Lichtenstein and Fischhoff (1980) and Griffin and Tversky (1992) argue that overconfidence can be attributed to the use of heuristics, which leads individuals to ignore vital information, which, in the view of these authors, is required to produce accurate probability estimates. Koriat et al. suggested that individuals are both biased in the retrieval of information (favouring positive evidence) and in their evaluation of the evidence (disregarding negative evidence); Griffin and Tversky suggested that individuals base their confidence on the strength of evidence available, and either ignore or under-utilize the weight of evidence. There is little empirical support for the Koriat et al. stage model, but some for the Griffin and Tversky strength and weight model.

Like Griffin and Tversky, Ferrell and McGoey (1980) are also pessimistic to the extent that they believe that information concerning predictive validity (such as discriminability and base-rate) is ignored, but imply that this can be corrected by the use of appropriate feedback. Unfortunately, the evidence that training will markedly improve calibration performance is weak (see below).

May (1986a, 1986b) is somewhat more optimistic, in that she argued that good calibration is expected if no "misleading" items are present, and the subjects are relying simply on their sensitivity to objective physical differences. However, Björkman, Juslin and Winman (1993) have shown that for true psychophysical judgements, underconfidence is observed, which they attribute to the fixed sensitivity of the sensory system and claim it is impossible to avoid. With respect to general-knowledge items, May claimed that with random sampling and the use of simple inference (a probabilistic syllogism) good calibration could be achieved, but that overconfidence was to be expected if the sampling procedure was biased. However, she believed that if items were misleading because subjects held "false knowledge" (such as a distorted cognitive map) then overconfidence would result.

Clearly the most optimistic theorists are those responsible for developing the ecological models. Both Gigerenzer, Hoffrage and Kleinbölting (1991) and Juslin (1993, 1994) believe that the miscalibration observed in general-

knowledge tests can be attributed to the biased sampling of stimulus items, and that with representative sampling individuals are well calibrated. Both models (and PMM theory in particular) provide the most complete explanation for the empirical findings (the Griffin and Tversky results notwithstanding), although to what extent these ideas can be extended beyond the general-knowledge domain remains to be seen.

Juslin believes that different mental representations are used in different stimulus domains, and neither internal cue theory nor PMM theory can provide a general explanation for the calibration of probabilities. Gigerenzer et al. have implied that PMM theory is general, and that it can explain calibration data in perceptual as well as cognitive tasks. However, the evidence for this proposal is not conclusive, and further research is required.

## 18.6 CAN ONE LEARN TO BE "WELL CALIBRATED"?

The attempts to improve individuals' calibration performance through the use of training with feedback have met with limited success. Ferrell (see Chapter 17) states that "... it is comforting that calibration can be improved relatively easily by suggestion and by training", but notes that "This optimism must be tempered by the findings that training ... does not seem to generalize very well" (page 430). Keren (1991) is even less optimistic; "The most disturbing finding obtained from training studies is that whatever modest improvement is achieved, it is hardly ever generalized to other tasks" (page 238). Other authors, however, put a very positive spin on the evidence for improvement through training. Russo and Schoemaker (1992) boldly state that "We believe that timely feedback and accountability can gradually reduce the bias toward overconfidence in almost all professions. *Being 'well calibrated' is a teachable, learnable skill*" (page 11, italics theirs). However, the evidence that outcome feedback alone is effective in reducing miscalibration is not encouraging (see Benson & Onkal, 1992 for a review).

The models we have reviewed vary in the amount of optimism they engender regarding the learning of good calibration. The most pessimistic models suggest that neither training nor experience will have an effect on calibration performance. This is either because miscalibration is a consequence of the manner in which the brain stores information (the memory trace model) or because the same heuristics—with the same limitations—are always used (the strength and weight model). Albert and Sponsler (1989) imply that someone with a poor memory will never be well calibrated, and that the miscalibration will be in the direction of overconfidence. Griffin and Tversky (1992) suggested that the bias in favour of the strength of evidence over its weight is incorrigible, and argued that calibration performance depends heavily on the



predictability of outcomes in a target domain. For example, they suggested that experts will be more overconfident than non-experts in an unpredictable domain (e.g. clinical psychology or the stock market) because they will give unwarranted credence to the validity of their expert knowledge.

The detection model is not entirely pessimistic because it allows for an improvement in calibration performance with outcome feedback, which allows the assessor to adjust his or her criteria on the evidence variable appropriately. However, this does not imply that an individual who becomes well calibrated in one domain will be well calibrated in another, and as we have noted, the evidence that outcome feedback improves calibration performance is very weak.

The expectations derived from the stage and process models are that certain types of training will lead to improvement in calibration under certain circumstances. For the stage model, training must be in the form of the generation of counter-arguments as described by Koriat, Lichtenstein and Fischhoff (1980), but again this seemed to have little effect on calibration performance. In the case of the process model we would anticipate that training in the form of the correction of false knowledge should lead to a reduction in overconfidence and thereby improve calibration.

Finally, we turn to the ecological models, which furnish quite specific predictions as to when training will, and will not have a beneficial effect on calibration, and provide a simple explanation for the poor results obtained when training with feedback has been examined. Training with outcome feedback *should* be effective when subjects are confronted with novel tasks, as this will allow them to learn the appropriate cues and cue validities required to make predictions. Any procedure which allows individuals to observe the covariation between variables in an ecologically valid setting should lead to an improvement in the quality of their subjective probability judgements.

Training with outcome feedback *will not* be effective with tasks such as standard general-knowledge tests, which contain an unrepresentative sample of items from the reference class. This prediction stems from the assumption that subjects will continue to use cues and report cue validities which are ecologically valid, but are not valid for non-representative stimuli. As Keren (1991) pointed out, most training investigations have used general-knowledge items, and we agree with him that the modest improvements which have been noted can be attributed to the fact that subjects receiving continuous feedback that their probability responses are too high will naturally lower them, but this is merely a "technical correction", and has nothing to do with improving probability judgements. This analysis also explains why any improvement does not generalize to other tasks. The quality of the calibration performance depends on the experience the subject has had with the target domain, and crucially, on how the stimuli used in the test have been selected. Thus, from

the perspective of the ecological models, there is cause for optimism with respect to training—but not for the notion of a general training in calibration.

## 18.7 CONCLUSIONS

We have argued that the ecological models, and PMM theory (Gigerenzer, Hoffrage & Kleinbölting, 1991) in particular, provide the most coherent account of how individuals realize subjective probability judgements, and afford the most satisfactory explanation of calibration performance with general-knowledge items.

The calibration of subjective probabilities has been studied in a variety of other task domains, and it remains an open question as to how successfully the ecological approach can be applied in other settings. For example, Björkman, Juslin & Winman (1993) have provided evidence that, when making psychophysical judgements, an alternative representation which leads to *underconfidence* is used. The representation is based on the subjective distance between the stimuli, and Björkman et al. argue that due to the fixed sensitivity of the sensory system, the bias cannot be avoided.[7] They also provide evidence that training has no effect on the underconfidence bias. However, in a recent paper Baranski and Petrusic (1994) have questioned the subjective distance model. In three experiments, these authors showed that it is possible to obtain *overconfidence* in psychophysical judgments when response accuracy is sufficiently reduced—either by putting the subjects under speed stress, or by reducing discriminability sufficiently under accuracy stress. They also studied decision time conditionalized on confidence category, and argued that their results were incompatible with both the subjective distance model *and* the detection model described earlier.

There are other task domains in which it may be implausible that PMM theory or internal cue theory applies in the form suggested by the ecological models. For example, weather forecasters are notoriously well calibrated (Murphy and Winkler, 1977; 1984) but it would seem unreasonable (but not impossible!) that a *single* cue is used to arrive at both a decision concerning precipitation and the associated probability. It is also difficult to see how the models can be applied to *episodic* memory tasks. What sort of interaction with the environment and encoding of the co-occurrences of events would help to decide that a stimulus item was present or absent during the encoding phase in a recognition memory experiment? Wagenaar (1988) showed that subjects were quite well calibrated (but demonstrated some overconfidence) for "old" items but extremely poorly so for "new" items in an old/new recognition test using words, syllables and numbers as stimuli. McClelland (1992) obtained similar results in a face recognition study. Wagenaar also showed that calibration was reasonably good when subjects were able to retrieve



information directly from episodic memory, but overconfidence became evident when they relied on inference rather than direct memory retrieval. We believe that in experiments of this type, probability judgements may be based on a form of representation not well captured by either PMM theory or internal cue theory in their present form. Further empirical and theoretical work is clearly needed.

With respect to a *general model* of calibration performance, Baranski and Petrusic (1994) have argued that the properties of decision times in calibration tasks place tight constraints on possible candidates (see also Wright & Ayton, 1988). Baranski and Petrusic suggest that an "appealing avenue for theoretical consideration" (page 426) might be a variant of Ferrell and McGoey's (1980) detection model—one which could account for the pattern of reaction-time and response probability relationships observed in their experiments. They favour an approach based on some form of evidence accumulator model (e.g. Link, 1992; Petrusic, 1992; Vickers 1970, 1979).

Recently, a colleague of ours who works in the area of judgement and decision making commented in a rather exasperated tone that it was about time that the calibration issue "was laid to rest". That outcome may still be some way off, but we feel confident that the days of "dust-bowl empiricism" are over, and that there is now a rich enough source of theoretical ideas to drive calibration research in a more productive direction.

## ACKNOWLEDGEMENTS

We would like to thank Peter Ayton, Nigel Harvey and A.R. Jonckheere for their advice and comments during the preparation of this chapter.

## NOTES

(1) The majority of calibration studies have used a 2AFC paradigm. For each item (e.g. "Absinthe is (a) a precious stone, (b) a liqueur") the subject selects one alternative and gives a probability rating that the choice is correct on a scale between 50% and 100%. In a full-range probability true task, the subject responds on a scale between 0% and 100% to indicate the degree to which they believe each statement to be true (e.g. "Absinthe is a precious stone") or confidence in the outcome of a future event (e.g. "What is the probability that it will rain tomorrow?"). For other task formats see Ferrell & McGoey (1980).

(2) In full-range tasks, base rate (the relative frequency of statements actually true, or of the occurrence of an event) has also been found to have an effect on calibration performance (see Smith & Ferrell, 1983). However the overestimation of the appropriate relative frequencies with base rates below 50%, and underestimation with base rates above 50%, has received far less attention than the overconfidence and hard–easy effects.

(3) Although information concerning the other alternative is not *explicitly* present

in the syllogism, it must presumably be the case that the individual knows that Hyderabad *is not* a capital city (see also, Gigerenzer, Hoffrage & Kleinbölting, 1991, page 523).

(4) Albert and Sponsler (1989) use the concept of "expertise" with respect to the accuracy of an assessor in making subjective probability judgements, and not to indicate the degree to which an individual is regarded as an expert within a particular knowledge domain.

(5) We assume this is a reference to the slight *underconfidence* that is often observed at the 50% point on the subjective probability scale in 2AFC tasks.

(6) However, Ferrell and McGoey (1980) noted that in many cases, the fit of the model was not so precise as to be statistically indistinguishable from the data.

(7) Whilst this may be true for *psychophysical* tasks, it does not follow that it is true for *all* perceptual tasks. Extracting information predictive of a target event from a complex and noisy stimulus display is likely to be a learnable skill, leading to improved calibration.

## REFERENCES


Albert, J.M. & Sponsler, G.C. (1989) Subjective probability calibration: A mathematical model. *Journal of Mathematical Psychology*, **33**, 298–308.

Baranski, J.V. & Petrusic, W.M. (1994) The calibration of confidence in perceptual judgments. *Perception and Psychophysics*, **55**, 412–28.

Benson, P.G. & Onkal, D. (1992) The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, **8**, 559–73.

Björkman, M. (in press) Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior and Human Decision Processes.*

Björkman, M., Juslin, P. & Winman, A. (1993) Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception and Psychophysics*, **54**, 75–81.

Brunswick, E. (1943) Organismic achievement and environmental probability. *Psychological Review*, **50**, 255–72.

Brunswick, E. (1955) Representative design and probabilistic theory in a functional psychology. *Psychological Review*, **62**, 193–217.

Dawes, R.M. (1980) Confidence in intellectual judgments. In E.D. Lantermann & H. Feger (eds.), *Similarity and Choice*. Hans Huber, Bern.

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

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

Fischhoff, B. & MacGregor, D. (1982) Subjective confidence in forecasts. *Journal of Forecasting*, **1**, 155–72.

Gigerenzer, G. (1991) How to make cognitive illusions disappear: Beyond "Heuristics and Biases". *European Review of Social Psychology*, **2**, 83–115.

Gigerenzer, G., Hoffrage, U. & Kleinbölting, H. (1991) Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, **98**, 506–28.

Griffin, D. & Tversky, A. (1992) The weighting of evidence and the determinants of confidence. *Cognitive Psychology*, **24**, 411–35.

Harvey, N. & Rawles, R. (1992) Probability matching in probabilistic mental models. *Bulletin of the Psychonomic Society*, **30**, 488.





Hasher, L. & Zacks, R.T. (1984) Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist*, **39**, 1372–88.

Jungermann, H. (1983) The two camps on rationality. In R.W. Scholz (ed.), *Decision Making under Uncertainty*. Elsevier, Amsterdam.

Juslin, P. (1993) An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology*, **5**, 55–71.

Juslin, P. (1994) The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organisational Behavior and Human Decision Processes*, **57**, 226–46.

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

Kahneman, D. & Tversky, A. (1982) On the study of statistical intuitions. In D. Kahneman, P. Slovic & A. Tversky (eds.), *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

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

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

Keren, G. (1991) Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, **77**, 217–73.

Koriat, A., Lichtenstein, S. & Fischhoff, B. (1980) Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, **6**, 107–18.

Lichtenstein, S. & Fischhoff, B. (1977) Do those who know more also know more about how much they know? The calibration of probability judgements. *Organizational Behavior and Human Performance*, **20**, 159–83.

Lichtenstein, S., Fischhoff, B & Phillips, L.D. (1982) Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (eds.), *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

Link, S.W. (1992) *The Wave Theory of Difference and Similarity*. Erlbaum, Hillsdale, NJ.

McBurney, D.H. (1990) *Experimental Psychology* (2nd edn.) Wadsworth, Belmont, CA.

McClelland, A.G.R. (1992) Facial identification: Is there a relationship between confidence and accuracy? *International Journal of Psychology*, **27**, 114.

McClelland, A.G.R., Bolger, F. & Tonks, E. (1993, January) The effects of discriminability and base rate on calibration of probabilities in a perceptual discrimination task. In N. Harvey & P. Ayton (Chairs). *Judgment and Decision Making*. Symposium conducted at the meeting of the Experimental Psychology Society London.

McClelland, A.G.R., Coulson, A.S. & Icke, S.E. (1990) Bias in meta-memory performance and its implications for models of memory structure. In J.-P. Caverni, J.-M. Fabre & M. Gonzalez (eds.), *Cognitive Biases*, North Holland, Amsterdam.

May, R.S. (1986a) Inferences, subjective probability and frequency of correct answers: A cognitive approach to the overconfidence phenomenon. In B. Brehmer, H. Jungermann, P. Lourens, & G. Sevo'n (eds.), *New Directions in Research on Decision Making*. North Holland, Amsterdam.

May, R.S. (1986b) Overconfidence as a result of incomplete and wrong knowledge. In R.W. Sholtz (ed.), *Current Issues in West German Decision Research*. Lang, Frankfurt.

Milburn, M.A. (1978) Sources of bias in the prediction of future events. *Organizational Behavior and Human Performance*, **21**, 17–26.

Murphy, A.H. & Winkler, R.L. (1977) Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest*, **2**, 2–9.
Murphy, A.H. & Winkler, R.L. (1984) Probability forecasting in meteorology. *Journal of the American Statistical Association*, **79**, 489–500.
Petrusic, W.M. (1992) Semantic congruity effects and theories of the comparison process. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 962–86.
Phillips, L.D. & Wright, G.N. (1977) Cultural differences in viewing uncertainty and assessing probabilities. In H. Jungermann & G. de Zeeuw (eds.), *Decision Making and Change in Human Affairs*. Reidel, Amsterdam.
Pitz, G.F. (1974) Subjective probability distributions for imperfectly known quantities. In L.W. Gregg (ed.), *Knowledge and Cognition*. Erlbaum, Hillsdale, NJ.
Poulton, E.C. (1989) *Bias in Quantifying Judgments*. Erlbaum, New York.
Ronis, D.L. & Yates, J.F. (1987) Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, **40**, 193–218.
Russo, J.E. & Schoemaker, P.J.H. (1992) Managing overconfidence. *Sloan Management Review*, Winter, 7–17.
Smith, M. & Ferrell, W.R. (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 Decisions*. North Holland, Amsterdam.
Staël von Holstein, C.S. (1971) An experiment in probabilistic weather forecasting. *Journal of Applied Meteorology*, **10**, 635–45.
Staël von Holstein, C.S. (1972) Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance*, **8**, 139–58.
Tversky, A. & Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science*, **185**, 1124–31.
Tversky, A. & Kahneman, D. (1983) Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, **90**, 293–315.
Vickers, D. (1970) Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, **13**, 37–58.
Vickers, D. (1979) *Decision Processes in Visual Perception*. Academic Press, New York.
von Winterfeldt, D. & Edwards, W. (1986) *Decision Analysis and Behavioral Research*. Cambridge University Press, New York.
Wagenaar, W.A. (1988) Calibration and the effects of knowledge and reconstruction in retrieval from memory. *Cognition*, **28**, 277–96.
Wright, G. (1982) Changes in the realism and distribution of probability assessments as a function of question type. *Acta Psychologica*, **52**, 165–74.
Wright, G. & Ayton, P. (1988) Decision time, subjective probability and task difficulty. *Memory & Cognition*, **16**, 176–85.
Wright, G. & Phillips, L.D. (1984) Decision making: Cognitive style or task-related behaviour? In H. Bonarius, G. van Heck, & N. Smid (eds.), *Personality Psychology in Europe*. Swets & Zeitlinger, Lisse.
Yates, J.F. (1982) External correspondence: Decomposition of the mean probability score. *Organizational Behavior and Human Performance*, **30**, 132–56.
Yates, J.F. and Curley, S.P. (1985) Conditional distribution analysis of probabilistic forecasts. *Journal of Forecasting*, **4**, 61–73.
Zakay, D. (1983) The relationship between the probability assessor and the outcomes of an event as a determiner of subjective probability. *Acta Psychologica*, **53**, 271–80.