9

Perspectives on Probability Judgment Calibration

Dale Griffin and Lyle Brenner

Introduction

"The Games could no more have a deficit than a man could have a baby!" So quoth Jean Drapeau, mayor of Montreal, shortly before the 1967 Montreal Summer Olympics ran a $2 billion deficit. Are such unrealistic pronouncements typical of human intuitive judgment? Such questions do not hold merely academic interest. For example, each reader of this chapter will probably be faced with choosing a course of action based on a physician's judgment about the outcome of a medical intervention.

Despite the importance of understanding the intuitive judgments of physicians, judges, and politicians, the study of probability calibration is primarily a laboratory-based enterprise, with theoretical controversies resolved (or not resolved) with reference to laboratory findings. There are two primary reasons for this apparently myopic focus. Jean Drapeau's quote illustrates one: expert pronouncements usually serve multiple functions beyond communicating the judge's own beliefs. Such judgments also serve to persuade, to inspire, and even to undermine opposing viewpoints. Second, the methods and paradigms for studying probability judgment were shaped by investigators who were interested not in the quality of likelihood judgments but in people's ability to monitor their own knowledge. Thus there are thousands of studies on confidence in trivia knowledge, compared to a handful of studies on the calibration of experts in the field (Koehler, Brenner, & Griffin, 2002).

Our review proceeds as follows. First, we review some general background necessary for understanding the theoretical controversies in the field. We then outline some classic findings as defined by the paradigm-setting review of Lichtenstein, Fischhoff, & Phillips (1982). We then describe five major classes of theories of judgmental calibration, and we examine how, and how well, these theories account for the dominant stylized facts in the literature. We conclude with a specific applied example and describe how the different theories account for the observed phenomena.

Calibration Curves: Graphical Displays of Calibration and Miscalibration

The quality of calibration can be assessed through *calibration curves* that represent a qualitative pattern of calibration, or through indices that summarize the degree of calibration quantitatively. We focus on graphical representations as they have had considerably more impact on conclusions in this field than summary indices (see, e.g., Yates, 1990).

Forced choice, half-range tasks

Laboratory studies of calibration have relied on a standard paradigm inherited from the cognitive psychologists studying metacognition, or one's knowledge about one's knowledge. A typical experiment consists of a subject answering many general knowledge or "almanac" questions (e.g. "Which is further north: Paris or New York?"), and then rating his/her confidence, in the form of a probability, that the chosen answer is correct. The questions are typically presented with 2 choices so that the possible probability ratings in the chosen alternative range from 50 per cent to 100 per cent.

Accuracy rates (Y) are plotted against confidence ratings (X) in a calibration curve. Overconfidence occurs when confidence exceeds accuracy; underconfidence occurs when accuracy exceeds confidence (see Figure 9.1). Mixed cases occur when a curve starts out on one side of the identity line (often above the line representing underconfidence for relatively low probabilities such as .5 or .6) and then crosses the identity line (typically below the line representing overconfidence for higher probabilities). In these cases, it is essential to examine the calibration curve in tandem with the response proportions at each level of expressed probability (Wallsten, 1996). The same mixed pattern may indicate aggregate overconfidence if most of the judgments are made with high confidence or aggregate underconfidence if most of the judgments are made with low confidence.

Full-range tasks

One common source of confusion (resulting from the predominance of half-range tasks) is that there are distinct patterns of judgment referred to by the label "overconfidence." When probabilities are assigned to a focal hypothesis on the full 0 to 1 probability scale, we can distinguish between two forms of overconfidence: *overprediction*, depicted by curve A in Figure 9.2, the tendency to assign probabilities that are consistently too *high*; and *overextremity*, depicted by curve C, the tendency to assign probabilities that are consistently too *extreme* (i.e., too close to either 0 or 1). In the case of binary hypotheses,

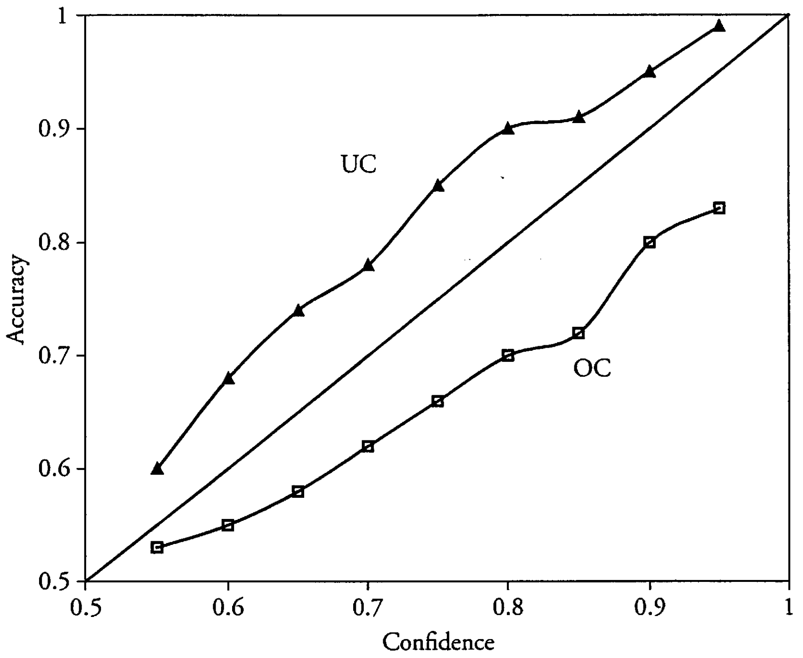

Figure 9.1 Sample half-range calibration curves

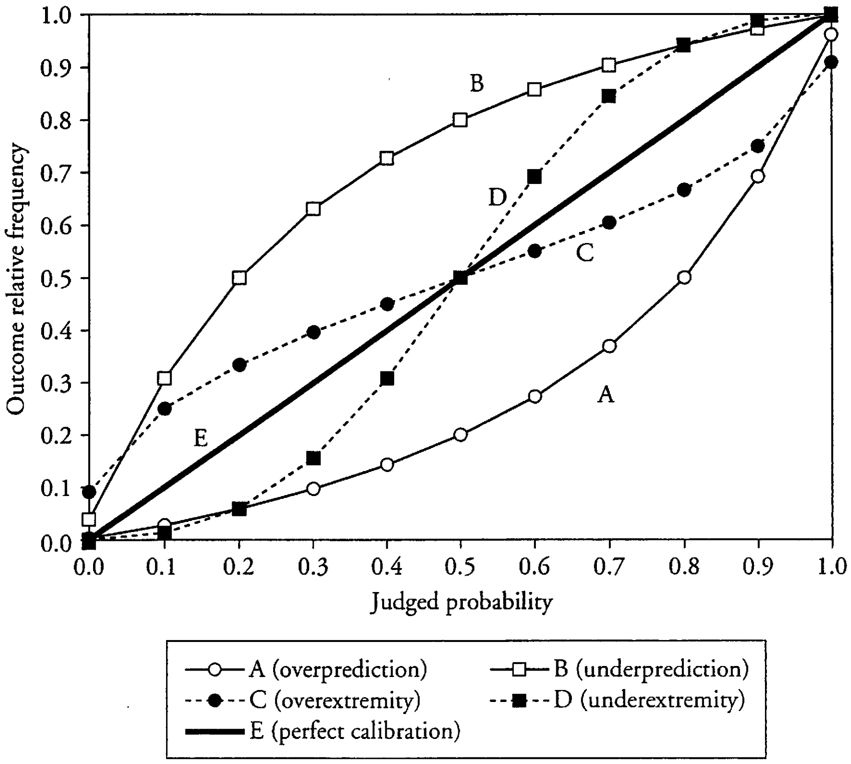

Figure 9.2 Sample full-range calibration curves

p.179

overextremity indicates an overestimation of whatever hypothesis the judge considers most likely. Thus, overconfidence, the poster child of judgmental biases, as a simple summary term does not uniquely identify one of these patterns (Wallsten & Budescu, 1983). Underestimation and underextremity can be defined similarly; underestimation (curve B) refers to assigning consistently too low probabilities to the focal hypothesis, and underextremity (curve D) refers to assigning probabilities that are not sufficiently extreme (i.e., probabilities too close to the middle of the scale.) Combinations of under- or overprediction and either of the extremity biases are also possible, and result in lines that cross the diagonal at points other than 50 percent. (See Harvey (1997) for a similar analysis.)

Liberman & Tversky (1993) called patterns of overextremity "generic overconfidence," and patterns of overprediction "specific overconfidence." Because overprediction refers to overconfidence in a specific designated hypothesis, it may be thought of as a bias towards that particular hypothesis. In contrast, in the case of binary hypotheses, overextremity indicates an overestimation of whatever hypothesis the judge considers most likely, and in that sense is independent of the focal hypothesis. Both overprediction and overextremity can be distinguished from optimistic overconfidence, which may be thought of as a specific form of overprediction – overestimation of the probability of events thought to be beneficial to the judge.

The Roots and Stylized Facts of Calibration Research

Early research on judgmental calibration was not aimed at discovering how people used probabilities, but in discovering how well people could assess or monitor their own knowledge. For example, Fullerton and Cattell (1892), Henmon (1911), and others all studied how well observers could introspect about whether their perceptions or college test answers were correct, and in particular whether observers could successfully report "partial knowledge." Henmon summarized his results as follows "While there is a positive correlation on the whole between degree of confidence and accuracy the degree of confidence is not a reliable index of accuracy" (pp. 200–1).

Two other parallel streams of early research were summarized by Lichtenstein et al. Research within meteorology on the accuracy of weather forecasts began very early in the twentieth century and unlike the psychological research, dealt exclusively with expert forecasters in the field. Research in the signal detection theory (SDT) paradigm studied the accuracy of confidence ratings in perceptual tasks during the 1950s and 1960s. The findings in these disparate fields were very similar: a preponderance of overconfidence, both in the overextremity and overprediction forms, with the degree of overconfidence depending on the difficulty of the task, and some scattered examples of underprediction. Lichtenstein et al. then reviewed scores of laboratory studies using almanac questions that showed the same pattern.

The Lichtenstein et al. review has been cited over 600 times, and usually for the following three points (in order of popularity): the predominance of overconfidence in the 2AFC almanac paradigm; the dependence of the degree of overconfidence on item difficulty; and the superb calibration of professional weather forecasters predicting rain in a Midwestern American city. In the manner of most secondary citations, the points are usually oversimplified compared to the comprehensive treatment in the review. The predominance of overconfidence was found across tasks, expertise, format, and method, across physicians and CIA operatives, weather forecasters and clinical psychologists. It is clearly not merely an artifact of the trivia or general knowledge paradigm. Furthermore, given that the amount of overconfidence is usually dramatic even with judgments of complete certainty, and with other forms of elicitation (e.g., odds, bets), the effect is not solely due to unfamiliarity with the probability scale or measurement artifacts due to the scale endpoints.

Similarly, the difficulty or "hard–easy" effect is not a hothouse phenomenon created by the clever concoction of a misleading set of general knowledge items. It has been found when participants differing in ability are compared, when participants with differing amounts of training are compared, and when item difficulty is defined by intrinsic qualities of the items rather than percent correct, as well as on post-hoc comparisons of high-accuracy versus low-accuracy items. In each case, the most difficult items or domains showed strong overconfidence, which declined and turned into underconfidence for the easiest items or domains. The same qualitative pattern was found in a signal detection study (Pollack and Decker, 1958), which examined the discriminability of words presented on earphones under conditions of high or low noise. An analogue to the difficulty effect was found when the proportions of "true" statements were manipulated in a one-alternative true-false task: overprediction when true statements were rare and underprediction when true statements were common.

The difficulty effect implies that there is a negative correlation between overall over/underprediction (Bias) and accuracy (Acc). It might appear that the difficulty effect is a statistical artifact, simply because the measure of overconfidence used contains the measure of difficulty: Bias = Conf − Acc. Let us examine this claim by calculating the covariance between Bias and Acc:

$$\text{Cov(Bias,Acc)} = \text{Cov(Conf,Acc)} - \text{Var(Acc)}$$

Note that this quantity certainly can be positive (contrary to the difficulty effect) if the correlation between Conf and Acc is large enough:

$$\text{Corr(Conf,Acc)} > \text{SD(Acc)} / \text{SD(Conf)}.$$

Thus, the difficulty effect is not a necessary feature of the method of data analysis, but is equivalent to a sufficiently low correlation between average confidence and accuracy across items.

From these "classic" and robust findings, we can summarize several "stylized facts" that theories of calibration need to explain.

1 Overconfidence is the predominant finding.
2 The degree of overconfidence depends on item difficulty (in the 2AFC case) and item base rate (in the full-range case). The calibration curve is relatively flat rather than rising with increasing probability.

3 Underconfidence is regularly found in very easy tasks and with very high base rates.
4 Excellent calibration is possible.

Theoretical Perspectives on Calibration

We now turn to five competing psychological accounts of probability calibration, and provide a conceptual framework for organizing the maze of empirical results. The five broad perspectives and their most important characteristics are summarized in Table 9.1. This set of theories is by no means exhaustive and we make no attempt to determine the "winner" of the theory competition, for our view is that these theories are like lenses that serve to organize the calibration data in different ways.

Optimistic overconfidence

The most influential perspective on miscalibration – at least for those outside the field itself – is the *optimistic overconfidence* perspective: people are notoriously subject to wishful thinking and self-enhancement, and thus provide probability estimates that are distorted by these self-serving motivations. This fits the dictionary definition of overconfidence: "The state or quality of being impudently or arrogantly self-confident." (*Roget's Thesaurus*, 1985). Biases thus reflect unwarranted arrogance or hubris, and overconfidence in the form of overprediction (curve A in Figure 9.2) should predominate and should vary according to the desirability of the outcome.

Conceptual background

The optimistic overconfidence perspective builds on several findings in the psychological literature: the *better than average effect* (e.g., Larwood & Whittaker, 1977, for managers; Svenson, 1981, for drivers; Alicke, 1985, for personal traits), the tendency to rate oneself as above the mean in positive skills and traits; *unrealistic optimism* (e.g., Weinstein, 1980), the tendency to rate oneself as more likely to experience positive events and less susceptible to negative events than others; *self-serving attributions* (e.g., Miller & Ross, 1975), the tendency to take credit for success and avoid blame for failure; and the *illusion of control* (e.g., Langer, 1975), the tendency to rate oneself as having some degree of control over random events. The account is furthered bolstered by the ubiquity of the *planning fallacy* (e.g., Buehler, Griffin, & Ross, 1994; Kahneman & Tversky, 1979), the tendency to believe that tasks will be completed more quickly and successfully in the future than they have been in the past, and *partisan belief polarization* (Hastorf & Cantril, 1954; Lord, Ross, & Lepper, 1979), the tendency for opposing partisans to interpret the same evidence as supporting their own divergent beliefs.

Conceptual critique

Both the generality of and the bases for self-enhancing forms of optimism have been questioned. Ironically, one of the papers most commonly cited to support the notion of

Table 9.1 Characteristics of theoretical models of calibration

	Optimistic overconfidence	*Confirmatory bias*	*Case-based judgment*	*Ecological models*	*Error models*
Primary assumption and emphasis	People are motivated to think well of themselves	People have information-processing bias to confirm hypotheses	Subjective probability is nonextensional and primarily focused on features of the case at hand (neglects relevant class-based evidence)	People adapt to and internalize statistical cues in natural environments; judged probabilities are unbiased estimates of ecological validity	Random error contaminates probability judgments; calibration of underlying beliefs is different than calibration of stated probabilities
Predicted biases	Overprediction	Overextremity	All patterns possible: Depends on information environment	Most patterns with nonrepresentative item selection; Good calibration with representative item selection	Some amount of overextremity is attributable to random error in responses; Overextremity and underextremity depending on method of data analysis
Expressed judgment represents	Hopes, wishes, self-enhancement	Proportion of evidence favoring most likely hypothesis	Evaluation of strength of evidence in case at hand	Reflection of ecological cue validity in natural environment	Internal "true-belief" probability plus random response error
Explanation for under-prediction and/or under-extremity	Underprediction: Depressive personality	None	Underprediction: High base rate Underextremity: High discriminability	Underprediction: Oversampling of high-probability items Underextremity: Sampling too many extreme-probability items	Underextremity: Items grouped by objective probability and inverse regression operating

Table 9.1 (*cont'd*)

	Optimistic overconfidence	*Confirmatory bias*	*Case-based judgment*	*Ecological models*	*Error models*
Explanation for predominance of overconfidence	Most people are optimistic	Confirmation is typical mode of reasoning	Forecasts are typically made for rare or difficult to predict outcomes	Studies of unrepresentative domains/items	Most predictions contain substantial random error
Explanation for overprediction and/or over-extremity	Overprediction: Focal hypothesis is positive/self-enhancing	Overprediction: Bias towards evidence for *focal* hypothesis Overextremity: Bias towards evidence for *more likely* hypothesis	Overprediction: Low base rate Overextremity: Low discriminability	Overprediction: Sampling too many low-probability items Overextremity: Sampling too many moderate-probability items	Overextremity: Large amount of random error in judgment
Explanation for the difficulty effect	None	More overlooked reasons for difficult items/domains	Changing discriminability and/or changing base rate of set neglected	Oversampling of difficult items varies from set to set	More random response error with more difficult judgments
Explanation for good calibration	No personal interest or motive	Even-handed consideration of alternative hypothesis	Moderate base rate and discriminability	Representative sample of items from natural environment	Little or no response error is necessary but not sufficient
Implications for debiasing	Reduce hubris; recruit arms-length judges	Explicit consideration of reasons for the alternative and against the focal	Training to increase attention to base rate and discriminability	Avoid nonrepresentative sampling; use frequency estimates when nonrepresentative	Reduce response error; Adjust for random error in analysis

self-serving attributions (Miller & Ross, 1975) advanced the claim that it is virtually impossible to separate motivational causes of self-serving attributions from informational causes. In particular, they note that people have much greater experience with success than with failure, and may thus explain them differently, even without any motivation to feel superior. In the same vein, recent commentators have noted that comparative optimism ("how do you compare to the average person"), one form of unrealistic optimism, may be due in part to an attentional bias and therefore less general than previously believed (Kruger, 1999). The common tendency to rate oneself as in the ninety-fifth percentile of drivers seems to be caused partly by an excessive focus on the self, with a corresponding lack of attention to the others serving as the basis of comparison. Thus, for domains where people have high absolute level of skill (e.g., driving) comparative optimism is found, but for domains where people have a low absolute level of skill (e.g., juggling), comparative pessimism is found, consistent with the argument that people anchor on their own level of skill and then adjust insufficiently for the comparative nature of the judgment (Kruger, 1999). (This interpretation does not hold for the many demonstrations of unrealistic optimism using "indirect" measures where individuals separately rate their own standing and the average person's standing.)

The planning fallacy, too, may be interpreted in informational terms: because the base rate of meeting predicted deadlines is relatively low, neglecting past experiences will give rise to apparently optimistic predictions. This interpretation is bolstered by the finding that the planning fallacy is equally pronounced in Japan and in Canada, despite the fact that the Japanese showed self-blaming attributions (Buehler, Otsubo, Lehman, Heine, & Griffin, 2003). Furthermore, the degree of optimism about future events is controlled by the temporal distance to the event, with events in the near future being regarded in a more evenhanded fashion (Gilovich, Kerr, & Medvec, 1993; Liberman & Trope, 1998; Shepperd, Ouellette, & Fernandez, 1996). All in all, these ambiguities and limitations in the optimistic bias account should lead to greater caution in its use as a *general* explanation of miscalibration.

Fitting stylized facts

The optimistic overconfidence account is only able to directly address the first stylized fact, the prevalence of overconfidence. For example, financial forecasts over the past century have been consistently over-optimistic (Hogarth & Makridakis, 1981). A survey of almost 3,000 new business owners revealed unrealistic optimism about their own business succeeding (81 percent probability of success for their own business vs 59 percent probability of businesses like theirs, whereas a realistic estimate is somewhere in the range of 30 to 70 percent, Cooper, Woo, & Dunkelberg, 1988). However, as noted above, these findings are open to interpretation in terms of information-based biases.

Confirmatory bias

A second broad perspective, largely eschewing motivational underpinnings, is the *confirmatory bias* perspective: People naturally search for evidence that supports their chosen hypothesis. Biases in calibration should thus reflect hypothesis-confirmation biases in attention, information gathering, and interpretation; consequently, overconfidence in the form of overextremity (curve C in Figure 9.2) should predominate.

Conceptual background

This account, too, is well supported by basic psychological evidence. When people test simple hypotheses about the relations between numbers and letters (e.g., Wason, 1968) or attempt to determine the personality type of an individual (e.g., Snyder & Swann, 1978) or even when teachers decide on a grade for a schoolchild (Rosenthal & Jacobson, 1966) they selectively search for confirmatory instances to "test" their theories.

Koriat, Lichtenstein, and Fischhoff (1980) argued that overconfidence arises in part from people's tendency to recruit evidence from memory that confirms the focal hypothesis. They offered a two-stage model in which the judge first selects her preferred option on the basis of a knowledge search, and then assesses her confidence by recruiting reasons supporting the preferred answer. The stronger and more numerous the reasons that are recruited, the greater is the confidence expressed in the selected answer. Because this process inclines the judge to overlook reasons *against* the selected answer, she is likely to be overconfident that the selected answer is correct.

Koriat et al. reported two results consistent with their model. First, asking subjects to generate reasons favoring and opposing both options reduced overconfidence relative to a control condition in which no such reasons were generated. Second, asking subjects to generate reasons contradicting their preferred alternative reduced overconfidence while generation of supporting reasons had no effect. Hoch (1985) also reported results consistent with the confirmatory search model in a study of predictions of graduating business school students regarding the outcome of their job searches. Asking students to generate reasons why the target event would *not* occur substantially reduced the observed overconfidence, whereas asking them to generate reasons supporting the target event's occurrence had no influence (see also Brenner, Koehler, & Tversky, 1996).

Conceptual critique

Considerable experimental evidence suggests that a confirmatory bias – like a tendency towards optimism – is responsible for creating *some* amount of overconfidence, particularly overextremity. However, direct evidence for its role and prevalence in probability judgments is scarce. The fact that providing reasons against a hypothesis can reduce overconfidence in general knowledge does not provide privileged support to the confirmatory bias account. Overconfidence may be created by any one of the mechanisms we discuss and still be reduced through strictures to "consider the opposite" (Lord, Lepper, & Preston, 1984). More studies are needed to distinguish the relative effects of hypotheses that are believed to be true and hypotheses that are wished to be true, as well as to distinguish confirmatory overconfidence from that caused by informational biases, such as neglecting the outcome base rate.

Fitting stylized facts

McKenzie (1997) offers a model that includes a parameter reflecting the extent to which evidence regarding alternative hypotheses is weighted in the confidence assessment, which ranges from 0 (complete neglect) to 1 (full weighting, equal to that placed on evidence regarding focal hypothesis). Anything less than a weight of 1 will produce overly extreme judgments. His model departs from the earlier confirmatory bias approaches in that he assumes that people have an unbiased sample of confirmatory evidence but simply neglect other evidence. A confirmatory bias model is used by Yates, Lee, Sieck, Choi, & Price (2002) to explain cultural differences in overconfidence. Such models can help to explain the prevalence of overconfidence reported in the literature using general knowledge questions. They are also, arguably, consistent with the observation of the difficulty effect. For easy tasks, there are likely to be proportionally few neglected reasons, since the majority of the evidence will point to the preferred (and correct) answer. On this account, we would expect substantial overconfidence for difficult tasks but not for easy tasks.

Confirmatory bias models cannot naturally accommodate the base rate effects found in full-range tasks or underextremity or underprediction, though underconfidence is commonly observed with easy tasks. Confirmatory bias may be one piece of the miscalibration puzzle, but it is not the whole story.

Case-based judgment

A third approach is the *case-based judgment* perspective, associated with the heuristics and biases and related literatures. From this perspective, judgment biases reflect the way that people intuitively perceive and assess relevant evidence. People focus on case-specific factors and neglect the information structure of the environment, leading to a pattern of miscalibration that includes all the curves drawn in Figure 9.2 (including the diagonal line of perfect calibration). The case-based perspective rests on the assumption that intuitive judgments of probability or likelihood are non-extensional; that is, that they are based on an evaluation of the individual case with little consideration of the set or class from which the case is drawn. Well-known findings in the heuristics and biases literature such as base rate neglect, non-regressive prediction, neglect of sample size, and the conjunction fallacy are demonstrations of the non-extensional nature of intuitive probability judgment (Kahneman, Slovic, & Tversky, 1982). This view is also consistent with much recent research indicating that judgments are often constructed based on internal sensations and cues (e.g., Koriat & Levy-Sadot, 1999).

Lichtenstein et al. (1982, pp. 316–17) provided an early description of the neglect perspective when they noted that "the hard–easy effect seems to arise from assessors' inability to appreciate how difficult or easy a task is." The neglect perspective was formalized in Ferrell & McGoey's (1980) Decision Variable Partition (DVP) model with cutoff parameters that were insensitive to changing evidence diagnosticity or outcome base rates. This model was successful in reproducing the difficulty effect in both the 2AFC case (where difficulty was neglected) and in the full-range case (where base rate was neglected). Griffin & Tversky (1992) noted the applicability of the heuristics and biases principles to the calibration context and proposed a *strength-weight* model of judged probability. According to this model, people intuitively focus on the strength of the evidence (how extreme is the evidence *in this case*) and then slightly adjust for the weight of the evidence (class-based factors such as sample size, base rate, and diagnosticity of the evidence). Such underadjustment leads to overconfidence when strength is high and weight is low and underconfidence when strength is low and weight is high; good calibration will generally occur when both are moderate. Furthermore, this model also unifies conservatism biases in belief updating (e.g., Phillips & Edwards, 1966) with the representativeness heuristic, as the underadjustment typical of conservatism is found with evidence of high diagnosticity (weight) but the overadjustment typical of the representativeness heuristic is found with evidence of low diagnosticity (weight).

Random Support Theory (RST; Brenner, 2003) supplements qualitative accounts such as the strength-weight model by characterizing the degree of case-based neglect in a given set of calibration data. RST, like Ferrell's DVP, uses a signal detection framework to link different outcomes to different confidence states in the judge; however, RST embeds the signal detection model within the broad non-extensional model of probability judgment provided by Support Theory (Tversky & Koehler, 1994; Rottenstreich & Tversky, 1997). An advantage of support-based models is that, in many cases, people can assess directly the extent to which the available evidence supports a given hypothesis. Koehler, Brenner, and Tversky (1997) report the results of a number of studies in which direct ratings of support are used successfully to fit probability judgments.

Conceptual critique

The empirical demonstrations used to underpin the heuristics and biases program have been the subject of many criticisms, ranging from claims that participants misunderstood the instructions to claims that the results might be restricted to paper and pencil tests of probabilistic reasoning. Each individual criticism may have some force with regard to a particular demonstration of a particular phenomenon. However, the large body of work is highly consistent and cannot be written off as a byproduct of experimental ingenuity or leading questions (Gilovich, Griffin & Kahneman, 2002). Furthermore, the calibration of experts in the field is consistent with the case-based model (Koehler et al., 2002, see Figure 9.3).

The chief difficulty with this class of models is that although people underweight class factors, they do use them to a degree that varies across situations. How does this happen? Does information about the weight of the evidence contaminate the assessment of its strength, without any attempt at Bayesian integration? Or is there something like an anchoring-and-adjustment mechanism that gives priority to the case-based evaluation but nonetheless consists of a separate evaluation for weight? These questions are critical issues for this account to address.

Fitting stylized facts

At the time of the Lichtenstein et al. (1982) review, the only existing model precise enough to be fit to empirical data was the decision variable partition model (DVP) of Ferrell and McGoey (1980). In the tradition of signal detection theory, this model describes confidence judgment as a process of partitioning an internal decision variable (which might be thought of as a feeling of confidence) into confidence categories that are used in making the overt judgment or response. Specifically, the model starts with the usual signal detection assumption that the decision variable can be represented using two unit-normal distributions, one for true or correct hypotheses and the other for false

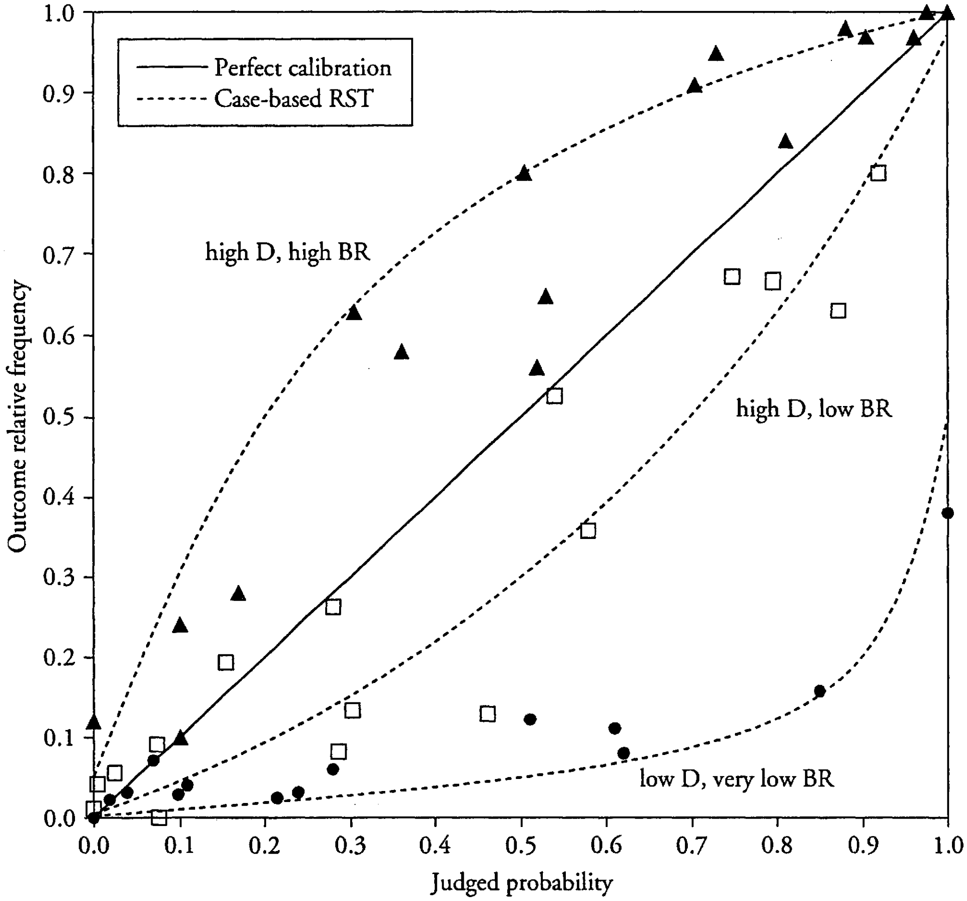

Figure 9.3 Calibration of physicians' probability judgments

or incorrect hypotheses. The former is assumed to have a higher mean than the latter, with the distance between them representing the discriminability of true and false hypotheses. The decision variable itself is not scaled in terms of probability; instead, the judgment is assumed to arise from a partition of the decision variable which assumes only that confidence is a monotonically increasing function of the decision variable.

The set of cutoff values established by the judge to create this partition is a crucial aspect of the partition model. Perhaps most impressive is the model's performance when supplemented by the assumption that the judge's set of cutoffs is insensitive to changes in task difficulty or base rate in the absence of performance feedback (Ferrell & McGoey, 1980; Smith & Ferrell, 1983; Suantak, Bolger, & Ferrell, 1996). Although there exists, for any given level of proportion correct, a set of cutoffs that would ensure perfect calibration (Gu & Wallsten, 2001), Ferrell and colleagues have found that the miscalibration observed in experimental contexts is often well accounted for by a single set of cutoffs that is not changed over large variations in the task environment. This insensitivity can produce any of the calibration patterns pictured in Figure 9.2, however, the model is agnostic about the nature of the underlying decision variable or where the cutoffs come from.

Brenner's RST model can fit the same range of data as the DVP model, but it incorporates a psychological theory of the determinants of confidence (support theory) and thus provides a more parsimonious and intuitive set of output parameters. The underlying dimension is now made up of two distributions of perceived *support*, for true and for false hypotheses. The distance between these two distributions is the discriminability parameter α. The set of cutoffs used in DVP is replaced by the focal bias parameter β (indicating sensitivity to base rate) and the extremity parameter σ (indicating sensitivity to discriminability). These parameters can be used to characterize almost any observed pattern of calibration in terms of the underlying process of support evaluation. However, highly specific predictions are made by constraining β and σ (usually fixing β to represent base rate neglect, and setting σ to near 1, indicating a moderate degree of variability in judged probabilities), while allowing discriminability and base rate to be free parameters of the environment.

Accounting for all the stylized facts then requires some additional assumptions. The prevalence of overconfidence implies that most judgment tasks that are studied (and probably most that are of interest in the real world) are difficult (leading to overextremity) and the outcomes of interest are rare (leading to overprediction). The difficulty effect implies that people tend not to alter the extremity by which they translate support into probability when the diagnosticity of evidence or the discriminability of the hypotheses change. Underprediction reflects those settings where the outcomes of interest are extremely common. Finally, there are some settings where the diagnosticity of the evidence is moderate and base rates are moderate – these settings should give rise to good calibration, even for untutored intuitive judgment. However, settings where good calibration is achieved in spite of extreme diagnosticity or extreme base rates require explanation beyond the basic theory (e.g., the calibrated prediction of rain in Chicago requires some explanation because the base rate of rain was moderately low).

Ecological probability

A fourth perspective is the *ecological probability* perspective; the key premise here is that people have highly accurate, adaptive knowledge of the probability of events in their natural environment. Because experiments rarely use stimuli that are representative of natural environments, studies find (or create) artificial biases in probabilistic judgment. Biases thus represent distortions induced by misleading empirical settings, and miscalibration should disappear when items are representative of the natural environment.

The above summary characterizes the second of two Brunswikian models that are relevant to calibration research. There is a long tradition of *lens model* approaches initiated by Brunswik himself (see Chapter 3, this volume). As Hammond noted (1998) "In short, ecological validity refers to the potential utility of various cues for organisms in their ecology (or natural habitat). Of course, the difference between the ecological validity of a cue and its actual use by an organism provides important information about the effective use of information by that organism." This is the central goal of the Brunswikian lens model social judgment theory approach: to determine what cues are used in judgment and how cue utilization compares to the ideal ecological validity of those cues (Hammond & Stewart, 2001). Most of these studies have focused on expert judgments, including many studies of meteorologists making probability judgments about weather events (e.g., Lusk, Stewart, Hammond, & Potts, 1990).

In a second wave of Brunswikian models of calibration, the focus on the use or misuse of ecologically valid cues by experts in the field has been replaced by the assumption of known ecological validities and by studies of students answering general knowledge questions. These models were motivated by May's (1986) observation that almanac studies finding overconfidence often used items hand-chosen to be challenging or even tricky, and her finding that judgments of overall accuracy ("how many did you get right") rarely showed the same degree of overconfidence. The first and probably best-known model of this type was the Probabilistic Mental Model (PMM) account developed by Gigerenzer, Hoffrage, and Kleinbölting (1991). A probabilistic model recruits a *reference class* from the natural environment (e.g., "all large cities in Germany"), and the reference class in turn recruits a series of cues. Confidence is determined by the cue validity, and "good calibration is to be expected if cue validities correspond to ecological validities" (Gigerenzer et al., 1991, p. 509). In general, items sampled from a well-defined reference class should meet this standard and show good calibration. If items are selected in a non-representative fashion, miscalibration will be observed.

In the first study testing these predictions, the calibration of a "representative set" of questions was compared with that of a "selected set" of general knowledge questions. The representative set was generated by randomly selecting 25 cities out of the 65 German cities with populations over 100,000. Participants judged pairs of these cities, decided which was larger, and indicated their confidence. After each block of 50 questions, participants estimated the number correct. The city-judgment task was substantially easier (72 percent) than the general knowledge task (52 percent), and in fact was much better calibrated overall, although judges showed substantial overextremity in the half-range task. Note that this finding is in accord with the difficulty effect as well as the selection effect. The authors attempt to address this ambiguity by selecting a portion of the city pairs that matched the difficulty of the general knowledge questions. As predicted by both the difficulty effect and the PMM, overconfidence in this selected set of difficult city questions showed substantial overconfidence. The same design in a second study led to a similar accuracy/difficulty confound (75 percent vs. 56 percent) and similar differences in calibration. Importantly, estimates of aggregate accuracy (frequency estimation over a set of problems) matched the observed accuracy rate of the difficult general knowledge questions and were substantially lower than the observed accuracy rate of the easier city questions. From these results, the authors concluded that overconfidence "disappeared" with representative sampling or with aggregate frequency judgments.

An extensive review of over 95 data sets (Juslin, Winman, & Olsson, 2000) found that the natural confounding of difficulty and representative versus selected data sets is almost complete. From the few studies that allowed the comparison at equal levels of difficulty, the overconfidence effect was much stronger for selected rather than representative sets (note, however, that regression artifacts make this comparison difficult to interpret). Although both overconfidence and underconfidence have been found with representative items, they rarely if ever show the extreme level of overconfidence found with difficult general item sets (possibly because the relatively easy representative sets allow less scope for overconfidence).

Conceptual critique

A central weakness of the second wave of ecological models is their continued focus on general knowledge and almanac questions, even ones randomly selected from world cities, countries, or death rates. Although it is important to question the role of difficult or tricky item sets (Keren, 1991), the research paradigm has lost the most important aspect of Brunswik's representative design: the actual environment and actual experts who have experience in selecting and using cues in that environment. Preliminary evidence from the limited number of calibration studies on experts in their natural environments reveals a dramatic pattern of miscalibration (Koehler et al., 2002) and should spur studies that examine cue use as well as ecological validity for representative judgments by lawyers, physicians, economic forecasters, and meteorologists. Furthermore, although it is both conceptually and practically difficult to define the appropriate reference class from which to sample items, representative design is satisfied by definition for the day-to-day judgments of experts.

The superior calibration of frequency judgments compared to probability judgments has found little support in subsequent studies. Instead, when frequency judgments are based on the same evidence as probability judgments, they show similar patterns of overconfidence (Brenner, Koehler, Liberman, & Tversky, 1996); otherwise, aggregate frequency judgments are simply lower than average confidence, leading to better overall calibration on difficult sets and poorer calibration on easy sets (Griffin & Buehler, 1999). Griffin and Tversky (1992) argued that the comparison of selected and representative sets should take into account not only the difficulty of the questions but also the strength of the impressions generated by the questions. Using random sampling from the same reference class (American states), they showed that holding accuracy constant at a low level, question sets that recruited strong impressions led to overconfidence while those that recruited weak impressions led to good overall calibration. Thus, neither representative samples nor specific levels of accuracy are sufficient to determine good calibration.

Fitting stylized facts

As described, the ecological models can account for overconfidence, the difficulty effect, and even underconfidence by invoking appropriately biased selection criteria. Good calibration should be associated with a representative sample. Although the key forms of miscalibration have been found even with representative sampling, the ecological models' assumption of an unbiased underlying representation of a true, ecological probability also has been incorporated in some of the error models that are reviewed below.

Error model

A fifth perspective is the *error model* or psychometric perspective. Error models attempt to separate the core underlying beliefs of the judge from the observed expressed probability judgments. In general, such approaches imply that uncorrected judgments will show overextremity biases, and that observed overextremity will be improved by correcting for or reducing the random response error.

Conceptual background

The psychometric justification for this approach is simple: when random error is added, the correlation between variables is attenuated. Erev, Wallsten, & Budescu (1994) proposed using the psychometric approach in the calibration domain by assuming that observed probability judgments reflect a systematic component (covert confidence) plus some amount of random error. Following this logic, even if the underlying true scores are unbiased, a significant amount of random error added at the response stage would lead to a lower correlation between judged probability and outcome than between true probability and outcome, and this, due to the effects of regression to the mean, would lead to observed overextremity.

The psychometric analogy also implies that the conclusions drawn by regressing or plotting Y (outcome frequency) as a function of X (judged probability) can be different from those drawn by regressing or plotting X on Y. If items can be classified by some objective probability, for example, general knowledge items can be classified by percent correct, then confidence can be plotted as a function of, or conditioned on, objective probability. If true underlying confidence is perfectly calibrated but random error is added, a regressive pattern is produced where high subjective probabilities are matched with lower outcomes (overconfidence in the high end of the scale) when the data are conditioned on judgment – but items with high objective probabilities are matched with lower subjective probabilities (underconfidence in the high end of the scale) when the data were conditioned on outcome. This pattern was labeled "simultaneous overconfidence and underconfidence" by Erev et al.

Conceptual critique

The results of Erev et al. can be interpreted in two ways, as a methodological prescription and as a descriptive model of probability judgment. The methodological prescription highlights the perils of diagnosing overconfidence on the basis of the calibration curve, since "error alone" can produce the appearance of overconfidence in such a curve even when underlying beliefs are unbiased. Brenner, Koehler, Liberman, and Tversky (1996) noted that the standard measure of overconfidence in 2AFC tasks, namely the difference between mean confidence and mean accuracy, provides an unbiased estimate of overconfidence which is not subject to the same kind of regression effect apparent in the calibration curve. Brenner (2000) questioned the logic of a model where observed overconfidence is relabeled based on assumptions about an unmeasured "latent" construct (see also Wallsten, Erev & Budescu, 2000).

Budescu, Wallsten, and Au (1997) assessed the relative contributions of random error and systematic bias (i.e., over- or underconfidence) to overall miscalibration. The reliability of probability judgments was assessed from replicate judgments and used to estimate the degree of miscalibration expected on the basis of error alone (i.e., in the absence of systematic bias), which was then used to construct a less strict standard of "ideal" performance than that which is usually employed, the identity line of perfect calibration.

(Klayman, Soll, González-Vallejo, & Barlas, 1999, offer another method for separating effects of systematic error and unreliability of judgments.) Using this method, Budescu et al. (1997) found substantial overextremity, even after correcting for the unreliability of the assessments, in a full-range task involving the relative populations of pairs of cities. As a descriptive model, then, the assumption of an unbiased "true score" subject to error is not a sufficient account of the miscalibration found in this and other laboratory tasks.

Error models are generally agnostic on whether the well-calibrated judge *should* take response error into account. Given that feedback from the environment should operate on observed judgments, one would expect learning to occur that would mitigate the effects of error by encouraging regressive adjustments to observed judgments. Clearly, patients would not be reassured upon learning that their miscalibrated physicians were suffering only from response error and their underlying probability assessments were perfectly calibrated (Brenner, 2000).

Fitting stylized facts

The error model approach as instantiated by Budescu et al. (1997) follows from Wallsten and González-Vallejo's (1994) stochastic judgment model, and is similar to Ferrell and McGoey's (1980) pioneering DVP signal detection model, invoking two normal distributions of covert confidence, one for true and the other for false statements. The underlying confidence measure is translated into stated probability by means of a set of cutoffs. The key innovation is in the modeling of within-state (i.e. within the true or false distributions) variance: Total within-state variance (σ^2) is composed of the variance between items within states (σ^2_b) and variance within items (σ^2_e). σ^2_e is interpreted as random error and is estimated by measuring multiple judgments of the same item (or "reversed" items, assuming binary complementarity). The probability assigned by the cut-off is perturbed by random error (on a log-odds scale).

Like Ferrell's DVP model, error models are sufficiently flexible in setting the cut-off levels so as to model any of the patterns presented in Figure 9.2. However, the psychometric approach is naturally designed to model overextremity. Error models are thus easily able to account for the prevalence of overconfidence, the tendency towards low confidence–accuracy correlations, and consequently the difficulty effect. However, it is not clear how error alone can produce any form of underconfidence.

Several researchers (Björkman, 1994; Juslin & Olsson, 1997; Juslin, Wennerholm, & Olsson, 1999; Soll, 1996) have recently offered modified ecological models in which stochastic error components have been introduced. In such models, the "internal" probability is only an estimate of the corresponding ecological probability, unbiased but subject to sampling error. Soll (1996), Juslin and Olsson (1997), and Budescu, Erev, and Wallsten (1997) have shown, using simulations, that a modified ecological model incorporating sampling error can produce overconfidence that increases with task difficulty.

One version of these models is able to account for underextremity in half-range judgments (summarized in Juslin et al., 1999). A key difference between this and other error models is that the perturbation takes place on the bounded probability scale rather than on the unbounded log-odds scale. Thus very easy tasks (the example used in the simulation was .95) are modeled by an underlying distribution producing many very extreme ecological probabilities, and in general added error will asymmetrically affect these probabilities so as to make the resulting judgments *underextreme* (as values such as .95 are limited in how much positive error can be added). Note, however, that all error models are designed to account for extremity biases; patterns of general bias such as overprediction and underprediction are not accommodated in such models.

Application and Example

Figure 9.3 presents calibration data from nine studies of practicing physicians' judgments about actual patients (adapted from Koehler et al., 2002). Each study was categorized in terms of the base rate of outcome (divided into high, moderate, and very low) and the physicians' ability to discriminate between cases when the event occurred and when it did not (moderately high or low). Each point on the graph represents a set of judgments and outcomes aggregated within a study by judged probability; a given study provided several data points. The data summarized in Figure 9.3 reveal that, across the different sets of medical events, physicians' probability judgments were sometimes too low (underprediction when base rate was high and discriminability was high), sometimes slightly too high (when base rate was low and discriminability was high), and sometimes much too high (overprediction when base rate was very low and discriminability was low).

It is instructive to consider how each of the five perspectives we have outlined might explain this pattern of data and would approach the problem of debiasing the physicians' judgments (see Chapter 16, this volume). The optimistic overconfidence perspective naturally leads to an expectation that overprediction would arise when outcomes were desirable and underprediction when outcomes were undesirable. However, this categorization does not account for the observed patterns in the data. The confirmatory bias perspective naturally leads to an expectation that more likely outcomes would be overpredicted and less likely outcomes would be underpredicted. In fact, the reverse is true. The case-based perspective naturally leads to an expectation that rare events will be overpredicted and common events will be underestimated (the dotted line refers to the predictions of RST assuming base rate and diagnosticity are completed neglected). This fits the obtained pattern, and leads to the suggestion that physicians' should be debiased with training on using set-based characteristics to overcome their case-based focus. The ecological perspective might suggest that, even though all judgments were made about real patients by expert physicians in their specific area of expertise, the categorization by base rate and discriminability still involves a selection effect. Averaged across all three groupings, the degree of bias is small and hence the ecological cues used by physicians may be unbiased. However, this approach offers little solace to the misclassified patients and no clear guidance as to how the categorized judgments may be debiased. Finally, the psychometric approach can explain the imperfect slopes of the lower two lines, but not the substantial vertical displacement, in terms of random error added at the response stage.

To the extent that there can be a "winner" in this competition among models, we believe the decision should be driven primarily by the philosophical and practical "fit" of the models to the problems we are trying to solve, rather than simply by the statistical goodness of fit of a model to experimental data.

References

Alicke, M. D. (1985) Global self-evaluation as determined by the desirability and controllability of trait adjectives, *Journal of Personality & Social Psychology*, 49, 1621–30.

Björkman, M. (1994) Internal cue theory: Calibration and resolution of confidence in general knowledge, *Organizational Behavior & Human Decision Processes*, 58, 386–405.

Brenner, L. (2000) Should observed overconfidence be dismissed as a statistical artifact? Critique of Erev, Wallsten, and Budescu (1994), *Psychological Review*, 107, 943–6.

Brenner, L. A. (2003) A random support model of the calibration of subjective probabilities, *Organizational Behavior & Human Decision Processes*, 90, 87–110.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996) Overconfidence in probability and frequency judgments: A critical examination, *Organizational Behavior & Human Decision Processes*, 65, 212–19.

Brenner, L. A., Koehler, D. J., & Tversky, A. (1996) On the evaluation of one-sided evidence, *Journal of Behavioral Decision Making*, 9, 59–70.

Budescu, D. V., Erev, I., & Wallsten, T. S. (1997) On the importance of random error in the study of probability judgment. Part I: New theoretical developments, *Journal of Behavioral Decision Making*, 10, 157–71.

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997) On the importance of random error in the study of probability judgment: Part II. Applying the stochastic judgment model to detect systematic trends, *Journal of Behavioral Decision Making*, 10, 173–88.

Buehler, R., Griffin, D., & Ross, M. (1994) Exploring the "planning fallacy": Why people underestimate their task completion times, *Journal of Personality & Social Psychology*, 67, 366–81.

Buehler, R., Otsubo, Y., Lehman, D. R., Heine, S. J., & Griffin, D. (2003) *Culture and Optimism: The Planning Fallacy in Japan and North America.* Manuscript submitted for publication.

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

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994) Simultaneous over- and underconfidence: The role of error in judgment processes, *Psychological Review*, 101, 519–27.

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

Fullerton, G. S. & Cattell, J. McK. (1892) On the perception of small differences: With special reference to the extent, force and time of movement, *Publications of the University of Pennsylvania: Philosophical Series*, no. 2.

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

Gilovich, T., Griffin, D. W., & Kahneman, D. (2002) *Heuristics and Biases: The Psychology of Intuitive Judgment.* New York: Cambridge University Press.

Gilovich, T., Kerr, M., & Medvec, V. H. (1993) Effect of temporal perspective on subjective confidence, *Journal of Personality & Social Psychology*, 64, 552–60.

Griffin, D. & Buehler, R. (1999) Frequency, probability, and prediction: Easy solutions to cognitive illusions? *Cognitive Psychology*, 38, 48–78.

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

Gu, H. & Wallsten, T. S. (2001) On setting response criteria for calibrated subjective probability estimates, *Journal of Mathematical Psychology*, 45, 551–63.

Hammond, K. (1998) Ecological Validity: Then and Now, online at: http://brunswik.org/notes/essay2.html.

Hammond, K. R. and Stewart, T. R. (eds.) (2001) *The Essential Brunswik: Beginnings, Explications, Applications.* New York: Oxford University Press.

Harvey, N. (1997) Confidence in judgment, *Trends in Cognitive Sciences*, 1, 78–82.

Hastorf, A. H. & Cantril, H. (1954) They saw a game; a case study, *Journal of Abnormal & Social Psychology*, 129–34.

Henmon, V. A. C. (1911) The relation of the time of a judgement to its accuracy, *Psychological Review*, 18, 186–201.

Hoch, S. J. (1985) Counterfactual reasoning and accuracy in predicting personal events, *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 11, 719–31.

Hogarth, R. & Makridakis, S. (1981) Forecasting and planning: An evaluation, *Management Science*, 27, 115–38.

Juslin, P. & Olsson, H. (1997) Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination, *Psychological Review*, 104, 344–66.

Juslin, P., Wennerholm, P., & Olsson, H. (1999) Format dependence in subjective probability calibration, *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 25, 1038–52.

Juslin, P., Winman, A., & Olsson, H. (2000) Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect, *Psychological Review*, 107, 384–96.

Kahneman, D. & Tversky, A. (1979) Intuitive prediction: Biases and corrective procedures, *TIMS Studies in Management Science*, 12, 313–27.

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

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

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999) Overconfidence: It depends on how, what, and whom you ask, *Organizational Behavior & Human Decision Processes*, 79, 216–47.

Koehler, D. J., Brenner, L., & Griffin, D. (2002) The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. Gilovich, D. Griffin, & D. Kahneman (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 686–715). Cambridge: Cambridge University Press.

Koehler, D. J., Brenner, L. A., & Tversky, A. (1997) The enhancement effect in probability judgment, *Journal of Behavioral Decision Making*, 10, 293–313.

Koriat, A. & Levy-Sadot, R. (1999) Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken and Y. Trope (eds.), *Dual-Process Theories in Social Psychology.* (pp. 483–502): Guilford.

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

Kruger, J. (1999) Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments, *Journal of Personality & Social Psychology*, 77, 221–32.

Langer, E. J. (1975) The illusion of control, *Journal of Personality & Social Psychology*, 32, 311–28.

Larwood, L. & Whittaker, W. (1977) Managerial myopia: Self-serving biases in organizational planning, *Journal of Applied Psychology*, 62, 194–8.

Liberman, N. & Trope, Y. (1998) The role of feasibility and desirability considerations in near and distant future decisions: A test of temporal construal theory, *Journal of Personality & Social Psychology*, 75, 5–18.

Liberman, V. & Tversky, A. (1993) On the evaluation of probability judgments: Calibration, resolution, and monotonicity, *Psychological Bulletin*, 114, 162–73.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982) Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 306–34). Cambridge: Cambridge University Press.

Lord, C. G., Lepper, M. R., & Preston, E. (1984) Considering the opposite: A corrective strategy for social judgment, *Journal of Personality & Social Psychology*, 47, 1231–43.

Lord, C. G., Ross, L., & Lepper, M. R. (1979) Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence, *Journal of Personality & Social Psychology*, 37, 2098–109.

Lusk, C. M., Stewart, T. R., Hammond, K. R. and Potts, R. J. (1990) Judgment and decision making in dynamic tasks: The case of forecasting the microburst, *Weather and Forecasting*, 5, 627–39.

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 A. Sevoaan (eds.), *New Directions in Research on Decision Making* (pp. 175–89). Amsterdam: North Holland.

McKenzie, C. R. M. (1997) Underweighting alternatives and overconfidence, *Organizational Behavior & Human Decision Processes*, 71, 141–60.

Miller, D. T. & Ross, M. (1975) Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, 82, 213–25.

Phillips, L. D. & Edwards, W. (1966) Conservatism in a simple probability inference task, *Journal of Experimental Psychology*, 346–54.

Pollack, I. & Decker, L. R. (1958) Confidence ratings, message reception, and the receiver operating characteristic, *Journal of the Acoustical Society of America*, 286–92.

Rosenthal, R. & Jacobson, L. (1966) Teachers' expectancies: Determinants of pupils' IQ gains, *Psychological Reports*, 115–18.

Rottenstreich, Y. & Tversky, A. (1997) Unpacking, repacking, and anchoring: Advances in support theory, *Psychological Review*, 104, 406–15.

Shepperd, J. A., Ouellette, J. A., & Fernandez, J. K. (1996) Abandoning unrealistic optimism: Performance estimates and the temporal proximity of self-relevant feedback, *Journal of Personality & Social Psychology*, 70, 844–55.

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, & A. Vari (eds.), *Analyzing and Aiding Decisions*. Amsterdam: North-Holland.

Snyder, M. & Swann, W. B. (1978) Behavioral confirmation in social interaction: From social perception to social reality, *Journal of Experimental Social Psychology*, 14, 148–62.

Soll, J. B. (1996) Determinants of overconfidence and miscalibration: The roles of random error and ecological structure, *Organizational Behavior & Human Decision Processes*, 65, 117–37.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996) The hard-easy effect in subjective probability calibration, *Organizational Behavior & Human Decision Processes*, 67, 201–21.

Svenson, O. (1981) Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, 47, 143–8.

Tversky, A. & Koehler, D. J. (1994) Support theory: A nonextensional representation of subjective probability, *Psychological Review*, 101, 547–67.

Wallsten, T. S. (1996) An analysis of judgment research analyses, *Organizational Behavior & Human Decision Processes*, 65, 220–26.

Wallsten, T. S. & Budescu, D. V. (1983) Encoding subjective probabilities: A psychological and psychometric review, *Management Science*, 29, 151–73.

Wallsten, T. S., Erev, I., & Budescu, D. V. (2000) The importance of theory: Response to Brenner (2000), *Psychological Review*, 107, 947–9.

Wallsten, T. S. & González-Vallejo, C. (1994) Statement verification: A stochastic model of judgment and response, *Psychological Review*, 101, 490–504.

Wason, P. C. (1968) Reasoning about a rule, *Quarterly Journal of Experimental Psychology*, 273–81.

Weinstein, N. D. (1980) Unrealistic optimism about future life events, *Journal of Personality & Social Psychology*, 39, 806–20.

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

Yates, J. F., Lee, J.-W., Sieck, W. R., Choi, I., & Price, P. C. (2002) Probability judgment across cultures. In T. Gilovich, D. Griffin, & D. Kahneman (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 271–91). Cambridge: Cambridge University Press.