Available online at www.sciencedirect.com
SCIENCE DIRECT®
Organizational Behavior and Human Decision Processes 97 (2005) 64–81
ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
www.elsevier.com/locate/obhdp

Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment ☆

Lyle Brenner [a,*], Dale Griffin [b], Derek J. Koehler [c]

[a] Warrington College of Business Administration, University of Florida, 212 Bryan Hall, Gainesville, FL 32611-7155, USA
[b] Sauder School of Business, University of British Columbia, Canada
[c] Department of Psychology, University of Waterloo, Canada

Received 16 June 2003
Available online 17 March 2005

Abstract

We describe four broad characterizations of subjective probability calibration (overconfidence, conservatism, ecologically perfect calibration, and case-based judgment) and show how Random Support Theory (RST) can serve as a tool for representing, evaluating, and discriminating between these perspectives. We present five studies of probability judgment in a simulated stock market setting and analyse the calibration data in terms of RST parameters. The observed pattern of calibration varies with the outcome base rate and cue value diagnosticity, as predicted by case-based judgment. A similar pattern of calibration is found in real-world judgments of experts in various domains. Case-based RST—defined as RST with stable parameter values—provides a parsimonious account of the substantial changes in calibration performance observed across different judgment environments.

Keywords: Subjective probability; Calibration; Support theory; Case-based judgment

Introduction

In modern society, both experts and laypeople are regularly faced with making probabilistic judgments about financial, medical, and personal outcomes. When are such probability judgments likely to be *calibrated*, and how might they be improved? These questions have fascinated decision researchers for decades, and continue to provoke controversy. After a brief review of several broad competing perspectives on the calibration of subjective probabilities, we introduce a general model of calibration that extends Support Theory (Fox & Tversky, 1998; Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994)—a general *coherence* model of subjective probability—to the question of the *correspondence* between stated probabilities and actual outcomes. We then present five studies demonstrating the use of this model, Random Support Theory (RST), both as a general theoretical framework and as a practical tool for characterizing and discriminating between different accounts of calibration. Finally, we compare the results of the laboratory studies with the calibration of experts in various domains making real-world judgments.

Divergent characterizations of calibration

Probability judgments of uncertain events are said to be perfectly calibrated if each set of events assigned a common probability judgment of p is in fact associated

☆ This research was supported in part by grants from the Social Sciences and Humanities Research Council of Canada and the Natural Sciences and Engineering Research Council of Canada. The article has benefited from the comments of the OBHDP associate editor and reviewers. We thank Michael Griffin and Carrie Charpentier for assistance in programming for the studies, and Baler Bilgin and Wouter Vanhouche for assistance in data collection.
* Corresponding author.
E-mail address: lbrenner@ufl.edu (L. Brenner).

0749-5978/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.obhdp.2005.02.002

with a corresponding relative frequency of p. We summarize four broad and conflicting characterizations about the calibration of probability judgment that have been offered in the literature (for integrative reviews and discussion see, e.g., Keren, 1991; Klayman, Soll, Gonzalez-Vallejo, & Barlas, 1999; Lichtenstein, Fischhoff, & Phillips, 1982; Yates, 1990). While it is generally recognized that no single perspective will fully account for all data, there are nevertheless general claims often made about the prototypical or central results in calibration studies. We present these four broad characterizations to summarize the most common themes.

Overconfidence

One characterization, consistent with the results of scores of laboratory calibration studies, is that people are generally overconfident. For example, when probabilities between 0.5 and 1.0 are assigned to an option chosen from a pair of alternatives (the so-called half-range paradigm), average judged probabilities (e.g., .75) are typically associated with accuracy rates that are considerably lower (e.g., .60). Overconfidence is seen by many as perhaps the prototypical summary finding within the calibration literature: "it is often believed that people's judgments are routinely overconfident" (Yates, 1990, p. 94), and "overconfidence is a reliable, reproducible finding" (Von Winterfeldt & Edwards, 1986, p. 539). The belief in overconfidence as a general feature of human judgment has spread well beyond the academic psychological literature; when describing investor behavior in the stock market, for example, Shiller (2000) commented that "some basic tendency toward overconfidence appears to be a robust human character trait" (p. 142).

At the process level, one common interpretation of the prevalence of overconfidence is that people tend to recruit evidence that confirms their focal hypothesis (e.g., the *confirmatory bias model* of Koriat, Lichtenstein, & Fischhoff, 1980). Another common interpretation is that people are generally *optimistic*, and thus tend to systematically overestimate the likelihoods of hoped-for or desirable events (e.g., Weinstein, 1980).

There are two distinct forms of overconfidence when probabilities are assigned to a focal hypothesis on the full 0 to 1 probability scale (Liberman & Tversky, 1993; Wallsten & Budescu, 1983): (a) *overestimation* (depicted in curve A in Fig. 1), the tendency to assign probabilities that are consistently too *high*; and (b) *overextremity* (depicted in curve C in Fig. 1), the tendency to assign probabilities that are consistently too *extreme* (i.e., too close to either 0 or 1). Consistent overestimation of the focal hypothesis is predicted by the confirmatory bias account of overconfidence; overestimation of desirable events and underestimation of undesirable events is predicted by the optimism account of overconfidence.

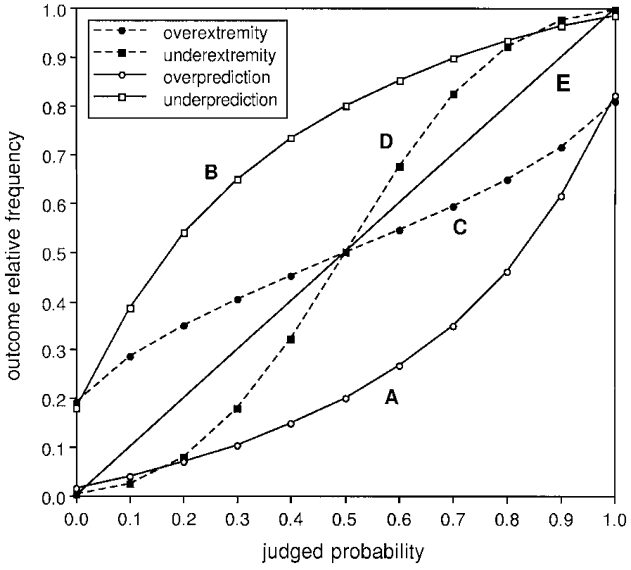

Fig. 1. Calibration curves illustrating distinct patterns of calibration and miscalibration (generated by RST simulations) *Note.* Five curves based on RST simulation assuming no focal bias ($\beta = 0$), and fixed judgmental extremity ($\sigma = 1$). Over/underprediction curves constructed assuming fixed discriminability ($\alpha = 1$) with varying target event base rate (BR = 20% for overestimation; BR = 80% for underestimation). Over/underextremity curves constructed assuming fixed base rate (BR = 50%) with varying discriminability ($\alpha = 0.5$ for overextremity; $\alpha = 2.0$ for underextremity).

Overextremity is consistent with the notion that people's beliefs do not sufficiently incorporate the uncertainty of their knowledge (Kahneman & Tversky, 1973). Insensitivity to such uncertainty also can account for the *difficulty effect* (Lichtenstein et al., 1982), the common finding that overconfidence is more substantial for difficult questions and is reduced (or reversed) with easier questions. Another interpretation of overextremity (and the difficulty effect) is that underlying beliefs may be perfectly calibrated but judgments are nevertheless too extreme because error is introduced in the response process (Erev, Wallsten, & Budescu, 1994; Soll, 1996; see also Brenner, 2000), or because judgment items are not representative of the domains to which people are adapted (Juslin, Winman, & Olsson, 2000).

Conservatism

A second and apparently contradictory characterization of calibration performance follows from classic studies investigating Bayesian updating (e.g., Phillips & Edwards, 1966). These studies indicated that people are overly *conservative* in that they are insufficiently sensitive to new diagnostic information. The phenomenon of conservatism implies *underextremity* (curve D in Fig. 1); people will overuse the middle of the probability scale (near .5) and underuse the extremes (near 0 and 1). A similar pattern of underextreme judgments also appears in studies of perceptual frequency estimation (e.g., Hollands & Dyre, 2000).

Erev et al. (1994) note that, depending on the method of data analysis, one may simultaneously observe both apparent overextremity and underextremity with the same set of data. For example, if objective probability (OP) is predicted from subjective probability (SP), as is typical in most calibration research, one may observe overextremity. With the very same data, however, if SP is predicted from OP—as when there are reliable normative standards of the correct probabilities of the judged events—one may observe underextremity. To avoid this potential confusion, throughout our discussion, we will always assume the traditional OP-as-a-function-of-SP analysis approach when characterizing patterns of calibration performance. Thus, the patterns in Fig. 1, and the terms overconfidence, underconfidence, overextremity, and underextremity are meant to depict distinct patterns of calibration performance that are not attributable to differences in the method of data analysis.

Good calibration through ecological adaptation

A third characterization of calibration performance is that, when judgment items are representatively drawn from a well-specified reference class, people's probability judgments tend to be well-calibrated (curve E in Fig. 1). The main premise of such *ecological* perspectives (e.g., Björkman, 1994; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994) is that through interaction with the environment, people internalize the associations between cues and events in the world and use this internalized knowledge when judging event probabilities. According to this view, laboratory observations of poor calibration are largely attributable to methodological artifacts such as the biased selection of judgment items. "If the set of general-knowledge tasks is randomly sampled from a natural environment, we expect overconfidence to be zero" (Gigerenzer et al., 1991, p. 512). Hybrid models in this tradition supplement the ecological approach with notions of error to describe how uncertainty due to sampling, as well as error in the sensory or response process can lead to observed overextremity despite well-calibrated underlying beliefs (Juslin & Olsson, 1997).

Case-based judgment

Finally, a fourth characterization is that, in making intuitive judgments of probability, people use simple mental operations (commonly called *heuristics*) that are sensitive to some features of the information environment, but insensitive to others (Kahneman & Tversky, 1973; Tversky & Kahneman, 1974), producing predictable patterns of calibration and miscalibration across different judgment environments (Griffin & Tversky, 1992; Massey & Wu, 2005). In particular, the case-based judgment perspective asserts that people's judgments rely primarily on evidence regarding the *particular case at hand* and tend to neglect relevant aggregate properties associated with the *class of instances* to which the case belongs. For example, a physician asked to assess a patient's risk of a certain medical condition will generally invoke (case-based) features of the patient, such as her health history and diet. In contrast, the physician is less likely to invoke (class-based) features that are merely seen as characteristic of the larger set of instances from which the case is drawn, such as the fact that the disease is very common nationwide.

In certain cases, aggregate characteristics may be considered as arguments in a case-based evaluation (Ajzen, 1977); for example, the base rate of a medical condition may be one argument considered by a physician ("there is a lot of that going around"). However, such usage typically will lead to underweighting of the class data compared to the ideal statistical model. Judgment may proceed by first forming a case-based impression, and then making a small, typically insufficient adjustment to account for class factors. Consistent with this view, Novemsky and Kronzon (1999) found that when base rates were used in a within-subjects design, they were used in an additive manner, rather than in the multiplicative manner required by the Bayesian model.

A model that incorporates both the effect of aggregate properties on the event outcome, and simultaneously the neglect of these properties by the judge, allows one to predict when probabilistic judgment will be appropriately calibrated, too high, too low, overly extreme, or insufficiently extreme (Griffin, Gonzalez, & Varey, 2000; Koehler, Brenner, & Griffin, 2002). Contingent on the features of the judgment environment, such a case-based model can predict the conditions under which each of the diverse patterns of judgment displayed in Fig. 1 are likely to be observed.

Each of the four alternative views described above can be characterized parsimoniously with Random Support Theory (RST), to which we now turn.

Support theory

Rather than attaching probabilities to events, Support Theory (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994) attaches subjective probabilities to descriptions of events, which are termed *hypotheses*. The construct of evidential *support* is introduced as an intermediary between hypotheses and judged probability. Each hypothesis A is assigned a support value $s(A) > 0$ which is interpreted as a measure of the strength of evidence for that hypothesis. The judged probability that hypothesis A rather than hypothesis B holds, assuming one and only one of them obtains, is given by

$$P(A,B) = \frac{s(A)}{s(A) + s(B)}. \quad (1)$$

In this representation, likelihood judgment reflects an assessment of the balance of evidence favoring the focal hypothesis rather than the alternative hypothesis. Previous work has investigated properties of this representation, and also properties of the support scale $s(\cdot)$ (Brenner & Koehler, 1999; Brenner & Rottenstreich, 1999; Fox, 1999; Koehler, 1996, 2000; Koehler, Brenner, & Tversky, 1997; Macchi, Osherson, & Krantz, 1999; Sloman, Rottenstreich, Wisniewski, Hadjichristidis, & Fox, 2004; for a review, see Brenner, Koehler, & Rottenstreich, 2002).

Support theory in its general form addresses the *coherence* of a set of probability judgments rather than their *correspondence* to the actual likelihood of outcomes as assessed in calibration analyses. To apply support theory to the study of calibration, Brenner (1995, 2003) developed a stochastic extension—Random Support Theory—that can model the calibration of subjective probabilities.

Random support theory

RST represents support as a random variable in order to reflect variability in perceived evidence strength across judgment items and judgment occasions. Similar to the approaches of Ferrell and McGoey (1980), Wallsten and Gonzáles-Vallejo (1994), and Budescu, Wallsten, and Au (1997), RST uses a signal-detection framework in which different distributions of support represent the strength of evidence for sets of correct and incorrect hypotheses. Unlike signal detection theory and these other models, however, RST does not invoke variable thresholds for converting the underlying random variable into a judgment; rather, support is mapped directly into a probability judgment based on the support theory representation in Eq. (1). This yields a parsimonious model in which the parameters of the underlying distributions of support can characterize distinct aspects of the judge's calibration performance. Furthermore, the underlying random variable is interpretable as support, which yields important psychological implications for the relationship between evidence judgment (which does not involve uncertainty) and probability judgment. We elaborate further on some of the similarities and differences between the related families of stochastic calibration models in the General Discussion.

Distributions of support in RST

We now describe the distributions of support needed to specify RST for full-range judgment tasks; Brenner (2003) addressed k-alternative tasks. Consider a judge estimating $P(A,B)$, the probability of focal hypothesis A relative to alternative hypothesis B. Suppose, as in the experiments described below, that hypothesis $A =$ "the price of Company X's stock will increase in the next financial quarter" and hypothesis $B =$ "the price of Company X's stock will decrease in the next financial quarter." RST specifies distributions for $s(A)$ and $s(B)$ under two circumstances: when A is in fact true and when B is in fact true. We will denote the actual outcome with a lower-case subscript; let A_a denote the (correct) "increase" hypothesis when the stock will actually increase, whereas B_a denotes the (incorrect) "decrease" hypothesis when the stock will actually increase. Similarly, A_b denotes the (incorrect) "increase" hypothesis when the stock in fact decreases, whereas B_b denotes the (correct) "decrease" hypothesis.

Support, representing the strength of evidence for a hypothesis, is necessarily non-negative. A convenient way of accommodating this constraint is to assume that the natural logarithm of support follows a normal distribution. RST is then easily described in terms of the distributions of the natural logarithm of support, rather than support itself, as follows (and displayed graphically in Fig. 2).

When A is correct:

$$\ln s(A_a) \text{ is normally distributed with mean } (\alpha+\beta)\sigma \text{ and standard deviation } \sigma. \tag{2a}$$

$$\ln s(B_a) \text{ is normally distributed with mean 0 and standard deviation } \sigma. \tag{2b}$$

When B is correct:

$$\ln s(A_b) \text{ is normally distributed with mean } \beta\sigma \text{ and standard deviation } \sigma. \tag{2c}$$

$$\ln s(B_b) \text{ is normally distributed with mean } \alpha\sigma \text{ and standard deviation } \sigma. \tag{2d}$$

Because support can be rescaled by an arbitrary constant multiplicative factor, log-support can be shifted by an additive constant without any loss of generality. Hence, the key feature in the specification of these distributions is the difference in the mean log-support for correct versus incorrect hypotheses. Using Eq. (1), the log-odds of the observable judged probability $P(A,B)$ can be represented as the log of the ratio of support values:

$$\ln\left(\frac{P(A,B)}{1-P(A,B)}\right) = \ln\left(\frac{\frac{s(A)}{s(A)+s(B)}}{\frac{s(B)}{s(A)+s(B)}}\right) = \ln\left(\frac{s(A)}{s(B)}\right). \tag{3}$$

Furthermore, the distributions of the log of the support ratios can then be determined from the distributions (2a) through (2d), assuming independence of $\ln s(A_i)$ and $\ln s(B_i)$:

$$\ln\left(\frac{s(A_a)}{s(B_a)}\right) \text{ is normally distributed with mean } (\beta+\alpha)\sigma \text{ and variance } 2\sigma^2. \tag{4a}$$

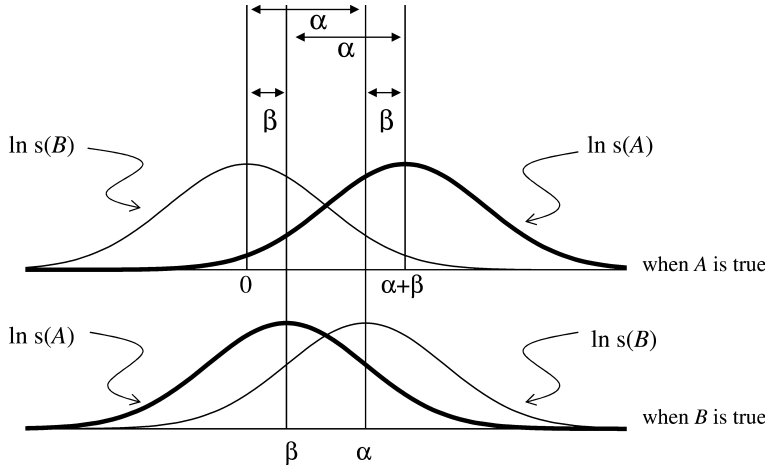

Fig. 2. Log-support distributions assumed by random support theory for focal hypothesis A (thick lines) and alternative hypothesis B (thin lines), conditioned on either A being true (top panel) or on B being true (bottom panel), illustrating role of RST parameters α and β. *Note.* Figure was constructed assuming $\sigma = 1$. Discriminability parameter α reflects the added support for a hypothesis when it is correct rather than incorrect; focal bias parameter β reflects added support for a hypothesis when it is the focal rather than the alternative hypothesis in the judgment.

$$\ln\left(\frac{s(A_b)}{s(B_b)}\right) \text{ is normally distributed with mean } (\beta - \alpha)\sigma \text{ and variance } 2\sigma^2. \quad (4b)$$

The assumption of independence requires no loss of generality in representing the judged probabilities, because any correlation (except $\rho = 1$) between the two log-support values can be absorbed into the variance parameter σ^2.

In summary, RST entails that the log-odds of judged probabilities will follow a particular distribution when the focal hypothesis A is correct (Eq. (4a)), and a separate distribution when the alternative hypothesis B is correct (Eq. (4b)). The parameters associated with these two distributions (α, β, and σ) can be estimated from the empirical conditional distributions of the judged probabilities.

Interpretations of parameters

The psychological interpretations of the RST parameters α, β, and σ follow from the specifications above in (2a) through (2d), and are displayed in Fig. 2. The *discriminability* parameter α is the additional support (in standardized log units) that is attached to the correct hypothesis relative to the incorrect hypothesis; hence it represents the judge's ability to view the correct hypothesis as being more strongly supported by the evidence. The parameter α is closely related to measures of discriminability from other analysis approaches. For example, when the base rate of the focal event is 50%, α is equal to d' from a signal detection analysis where the judge simply determines which of the two hypothesis is more likely. The discriminability parameter α is also a linear transformation of the ordinal correlation measure of discriminability suggested by Liberman and Tversky (1993).

The *focal-bias* parameter β is the additional support (again, in standardized log units) that is attached to the focal hypothesis A, regardless of whether it is correct or not. Positive values of β indicate a tendency to give systematically larger judgments to the focal hypothesis. As will be seen below, the appropriate value of β for good calibration will depend on the base-rate of the focal event. As such, a non-zero value for β does not necessarily indicate bias in the sense of an error or inaccuracy; it simply indicates a systematic shift in support for the focal hypothesis, regardless of the hypothesis's occurrence or non-occurrence.

Note also that the interpretation of β depends closely on the nature of the events that are judged. If a set of judgments always entails the same focal hypothesis (e.g., a meteorologist judging the probability of rain, or an economist judging the probability of recession), then β is interpretable as the overall additional support towards that particular hypothesis. Given a consistent focal hypothesis for a set of judgments, β does *not* necessarily indicate an overall bias towards hypotheses in the focal *position*; rather it indicates a bias towards the specific hypothesis that is repeatedly evaluated as the focal hypothesis in that judgment task. Therefore, the presence of a non-zero β in a task with a common focal hypothesis does not necessarily imply violations of binary complementarity in probability judgment, one of the assumptions of support theory (cf. Macchi et al., 1999; Brenner & Rottenstreich, 1999); indeed, the representation of judged probability as balance of support (Eq. (1)) that RST takes from support theory implies binary complementarity. In alternative representations which do not assume binary complementarity, an overall shift in support for whatever appears as the focal hypothesis could be reflected in β.

The *extremity* parameter σ is the standard deviation of log-support for each of the four hypotheses A_a, A_b, B_a, and B_b. This parameter represents extremity of judgment; as the variability of log-support increases, the variability of judged probability also increases, and the judge tends to use the extremes of the probability scale more often. The intuition for this interpretation is that larger values of σ imply a greater likelihood that the support for the focal hypothesis and the support for the alternative hypothesis will be highly divergent, and therefore yield a probability judgment near the extremes of 0 or 1.

The final parameter needed to specify the model is the outcome base rate, denoted BR, which is an exogenous feature of the judgment environment.

Illustrating RST

We illustrate the interpretation of the RST parameters by fitting the model to a well-known data set from Keren (1987), who studied calibration in amateur and expert bridge players' judgments of the probability of successfully fulfilling bridge contracts. Using the group-level mean calibration data from Keren (1987), we determined the optimal RST parameter estimates (to minimize weighted squared deviations between predicted and actual performance, for each of the two groups of judges). The resulting RST parameter estimates demonstrate psychologically meaningful differences between the prediction performance of amateurs and experts. In terms of discriminating between contracts that would succeed and those that would fail, experts ($\alpha = 0.96$) were substantially better than amateurs ($\alpha = 0.59$). Both groups were about equally successful in achieving the proposed contracts (experts 55%, and amateurs 60%), presumably because the experts were more ambitious and aggressive in their bidding and attempted to achieve more difficult contracts. Despite their similar success rates, amateurs were far more optimistic ($\beta = 0.95$) about the success of contracts than were experts ($\beta = -0.05$), who showed essentially no bias towards the focal hypothesis. Finally, amateurs were substantially more extreme in their probability judgments ($\sigma = 1.34$) than were experts ($\sigma = 0.90$). Equivalently, amateurs were relatively more confident in whatever they judged to be most likely (either success or failure of the contract). Ferrell (1994) provides an analysis of the Keren (1987) study using the decision-variable partition model of calibration, yielding similar conclusions, but in terms of the cutoff parameters that characterize that model of calibration.

Values of RST parameters required for Bayesian judgment

Given a particular outcome base-rate and level of discriminability α, there exist values of $\beta = \beta^*$ and $\sigma = \sigma^*$ within RST that will produce Bayesian and hence perfectly calibrated judgments. Algebraically, these optimal values can be determined by matching the judged probability predicted by the model to the objective Bayesian probability of the outcome derived from the support distributions in (4a) and (4b). Consider an arbitrary judged probability $P(A,B) = \frac{1}{1+\exp(-j)}$ so that the judged log-odds are expressible as $j = \ln\left(\frac{P(A,B)}{1-P(A,B)}\right)$. Bayes's rule can be used to determine the actual log-odds of A being correct conditional on the judged log-odds j. Setting the actual log-odds equal to the judged log-odds allows us to solve for the optimal RST parameters σ^* and β^* that imply perfect Bayesian judgment. As shown in Appendix A, these optimal RST parameters are:

$$\sigma^* = \alpha;\ \beta^* = \frac{1}{\alpha}\ln\left(\frac{\mathrm{BR}}{1-\mathrm{BR}}\right). \tag{5}$$

Thus, for perfect calibration the optimal extremity parameter σ^* must precisely follow (indeed, exactly equal) the discriminability parameter α, and the optimal bias parameter β^* must follow the outcome base rate (transformed to log-odds and also scaled by α). Appendix A provides the full derivation and some intuitions for these results.

Using RST to model alternative characterizations of calibration

Having identified the optimal β^* and σ^* for perfect calibration, the four characterizations of calibration described previously can now be expressed in terms of RST. Table 1 depicts these four accounts in the form of constraints on the RST parameters. Overconfidence in the form of *overestimation* (curve A in Fig. 1) implies that β is consistently greater than the optimal β^*. Overconfidence in the form of *overextremity* (curve C) implies that σ is systematically greater than its optimal value, $\sigma^* = \alpha$. In contrast, *underextremity* associated with conservatism (curve D) implies $\sigma < \alpha$. Under conditions of representative sampling of judgment items, ecological models imply perfect calibration (the diagonal line E), reflected in RST by $\beta = \beta^*$ and $\sigma = \sigma^*$.

Finally, the case-based judgment account implies that support (and hence judged probability) primarily reflects

Table 1
Predicted RST parameter values for alternate perspectives on probability calibration

1. Overconfidence	
a. Overestimation	$\beta > \beta^*$ in all conditions
b. Overextremity	$\sigma > \sigma^*$ in all conditions
2. Conservatism (Underextremity)	$\sigma < \sigma^*$ in all conditions
3. Ecological Rationality (Good calibration)	$\beta = \beta^*$ and $\sigma = \sigma^*$ under representative sampling
4. Case-based judgment	$\frac{\partial \beta}{\partial BR} < \frac{\partial \beta^*}{\partial BR}$ and $\frac{\partial \sigma}{\partial \alpha} < \left(\frac{\partial \sigma^*}{\partial \alpha} = 1\right)$ across all conditions.

the evidence related to the particular case at hand, and thus is generally insensitive to aggregate statistical properties of the set of judgment items such as outcome base rate and discriminability as determined by the diagnosticity of the available evidence. According to this account, the parameters of the RST model that normatively "ought" to reflect the judge's internalization of these aggregate properties for good calibration (β and σ) will instead remain roughly constant, despite changes in environmental features that influence the outcome base rate (BR) or evidence diagnosticity (α). As a result, depending on the specific levels of evidence diagnosticity and base rate, the case-based judgment model predicts good calibration (E) when both base rate and diagnosticity are moderate; overprediction (A) when base rate is low; overextremity (C) when diagnosticity is low; underprediction (B) when base rate is high; and underextremity (D) when diagnosticity is high.

Indeed, the particular patterns of miscalibration shown in Fig. 1 were generated by RST by holding constant the values of the parameters β and σ while varying the values of the outcome base rate BR and discriminability α. In all cases, the parameters were fixed at $\beta = 0$ and $\sigma = 1$. Overprediction (curve A, with $\beta > \beta^*$) was generated by a low base rate (BR = 20%) and underprediction (curve B, with $\beta < \beta^*$) by a high base rate (BR = 80%), both with moderate discriminability ($\alpha = 1$) that was appropriately matched by σ (i.e., $\sigma = \sigma^* = 1$). Overextremity (curve C, with $\sigma > \sigma^*$) was generated by a low value of discriminability ($\alpha = 0.5$), and underextremity (curve D, with $\sigma < \sigma^*$) by a high value of discriminability ($\alpha = 2.0$), both with moderate base rate (BR = 50%) that was appropriately matched by β (i.e., $\beta = \beta^* = 0$).

In the studies that follow, we examine calibration data from a simulated stock market prediction task in order to discriminate between the four different characterizations of calibration. We fit RST to the individual-level data and compare the resulting parameter estimates across experimental manipulations of base rate and evidence diagnosticity to test the predictions of the four characterizations of calibration listed in Table 1.

Overview of stock market studies

We present five studies conducted in a web-based simulated stock market setting in which participants predicted the direction of stock price changes for a series of companies. For each company, participants received case-specific sales and cost information. High sales and low costs were associated with increases in stock prices, and low sales and high costs were associated with decreases in stock prices, as one might expect. The magnitude of this association (i.e., the diagnostic value of the cues), however, was learned from experience in the simulated stock market environment, as was the overall prevalence (i.e., base rate) of stock price increases in the market.

Each study began with a training session in which participants received outcome feedback after making binary predictions for a series of companies. For each company, the cue values were presented as colored bars in a chart and participants were asked to predict whether the company's stock price would increase or decrease in the next financial quarter. After entering a prediction, participants were informed whether the company's stock price had actually increased or decreased. Following the training session, participants completed the probability judgment task that is the focus of our analyses. Participants judged a new series of companies, again accompanied by sales and cost cue values, and were asked to assess the probability that the company's stock price would increase the following quarter. Responses were made on a 0–100% sliding probability scale, separated into intervals of 5%. Participants were paid based on the accuracy of their judgments using an incentive-compatible payoff scheme based on the Brier score of their judgments; as a proper scoring rule, the Brier score encourages honest reporting of subjective probabilities. A lottery was used to distribute bonus prizes to a random subset of participants; bonuses ranged up to $60.

In both the training and judgment sessions, the diagnostic value of the cues and the overall base rate of increasing stock prices were experimentally manipulated. In Studies 1, 2, and 5, these independent variables were varied between subjects, with outcome feedback in the judgment trials omitted in Study 1 and provided in Studies 2 and 5. In Studies 3 and 4, either cue diagnosticity or base rate was varied within subjects; in Study 3, the within-subject factor was blocked whereas in Study 4 it varied from trial to trial.

Cue diagnosticity was varied as follows. Each cue value (domestic sales and costs, and foreign sales and costs), conditioned on whether it was associated with a stock price increase or decrease, was represented as a normally distributed variable with unit variance. The sales cue distributions for companies with increasing stock prices had a greater mean than those for companies with decreasing stock prices, and vice versa for the costs cue distributions. The degree of separation between cue distributions for companies with increasing and decreasing stock prices determines the diagnostic value of each cue. For both sales and costs, the diagnostic value of domestic indicator cues was set to be higher than that of foreign indicator cues. Details of the diagnostic value of the cues are presented in the method sections of the individual studies.

The base rate of stock price increases (i.e., overall "bullishness" or "bearishness" of the market) was also varied in Studies 1–4. In the low BR condition, 40% of

companies had stock price increases, whereas in the high BR condition, 70% of companies had stock price increases. In Study 5 only diagnosticity was manipulated, and BR was held constant at 50%.

The set of companies associated with a particular condition was constructed by first setting the proportion of companies with increasing stock prices according to the desired base rate, and then sampling cue values for each company from the appropriate distribution (depending on diagnosticity condition and whether the stock price was to increase or decrease the next quarter). One subset of trials constructed in this manner was used in the training session, and a separate subset was used in the judgment trials. Cue values were the same for subjects within each condition (but of course differed across conditions).

Note that participants were exposed to unbiased random samples from the relevant population in both training and judgment trials. Because participants were able to directly experience the base rate of company success and the diagnostic value of the cues, through exposure to representative samples of items, this design avoids the common objections to "scenario studies" of probability judgment (e.g., that items are selected in a nonrepresentative manner, or that conversational norms or ambiguous language may lead participants to misinterpret, ignore, or simply disbelieve the stated base rate).

Studies 1 and 2

Because of their similar designs (and similar results), we report the results of the first two studies together.

Method

Participants

Participants were 176 business students at the University of British Columbia and the University of Florida. Three participants were dropped from the analysis because they did not use the cues appropriately, as indicated by outlying negative ($n = 2$) or near-zero ($r < .15$, $n = 1$) correlations between judged probability and outcome in the judgment trials.

Design and procedure

Both base rate (BR) and diagnosticity (D) were varied between subjects, yielding four experimental groups based on the procedure described in Overview of Stock Market Studies. In both studies, the distribution of participants across the four conditions was approximately equal: In Study 1 (no feedback during judgment trials), the sample sizes ranged from 18 to 22; in Study 2 (feedback during judgment trials) the sample sizes ranged from 21 to 26. Each participant completed 60 training trials, followed by 40 judgment trials in which they assessed the probability that a designated company's stock price would increase in the next financial quarter, p(increase).

The proportion of trials with increasing stocks was 40% in the low BR condition and 70% in the high BR condition. Level of diagnosticity was manipulated between subjects as follows. In the low diagnosticity (low D) condition, the separation between increasing stock and decreasing stock cue distributions was 0.8 standard deviations (*SD*s) for domestic indicator cues and 0.4 *SD*s for foreign indicator cues. In the high diagnosticity (high D) condition, the separation was larger: 1.2 *SD*s for domestic indicator cues and 0.8 *SD*s for foreign indicator cues. If a simple "company performance" measure is generated from the four cues by adding total sales and subtracting total costs, the resulting value correlates 0.50 with the dichotomous outcome variable in the low D conditions and 0.68 in the high D conditions.

Results

Calibration of probability judgments

The left-hand panel of Fig. 3 displays the group calibration curves for each experimental condition. In all studies, probability judgments were grouped into 7 intervals (0–10%, 15–25%, 30–40%, 45–55%, 60–70%, 75–85%, and 90–100%) to produce smoother calibration curves. Data from Studies 1 and 2 have been combined in this figure to save space, and the same general pattern of results holds in both studies; in comparisons of the average values of σ and β across comparable conditions in the studies, there were no significant differences (Fs < 2.5; also see Table 2). The primary result is that the shape of the calibration curves is visibly influenced by both the base rate and diagnosticity manipulations. The base rate manipulation influenced the *elevation* of the calibration curves; for both levels of diagnosticity, the high BR calibration curves are located above the low BR calibration curves. Companies assigned the exact same judged p(increase) probability were more likely to be associated with an actual increase in stock price in the high BR than in the low BR condition. The diagnosticity manipulation, in contrast, influenced the *slope* of the calibration curves; the high D calibration curves are steeper than the low D calibration curves. Changes in judged p(increase) probabilities are associated with larger actual changes in the likelihood of a stock price increase in the high D than in the low D conditions.

RST model fit to data

The RST model was fit to each participant's data, yielding individual estimates of σ, α, and β for each participant. The individual-level parameters were estimated using a method of moments approach. The means and variances of each participant's two conditional log-odds distributions were computed, and the values of the parameters

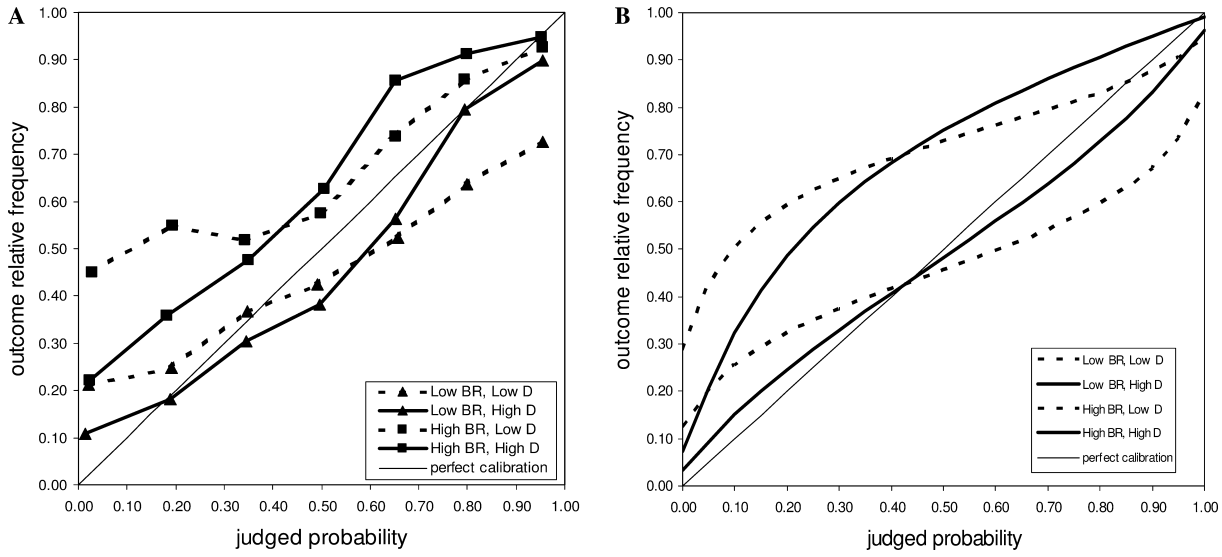

Fig. 3. Observed calibration curves (A) and curves predicted by case-based RST (B) for each combination of outcome base rate (BR) and evidence diagnosticity (D) in Studies 1 and 2. *Note.* Low BR = low base rate (40%); High BR = high base rate (70%); Low D = low diagnosticity ($\alpha = 0.62$); and High D = high diagnosticity ($\alpha = 1.17$). (B) RST parameter values and σ are held constant across conditions, $\beta = -0.21$, $\sigma = 1.47$.

Table 2
Mean individual RST parameter estimates (and Standard Errors) for Studies 1 and 2, by diagnosticity and base rate conditions

	Low BR (40%)	High BR (70%)
Study 1 (no judgment feedback)		
Low D		
σ	1.31 (.07)	1.10 (.05)
$\alpha = \sigma^*$	0.73 (.05)	0.63 (.04)
β	−0.31 (.15)	−0.05 (.14)
β^*	−0.46 (.04)	1.50 (.14)
High D		
σ	1.11 (.05)	1.12 (.04)
$\alpha = \sigma^*$	1.30 (.05)	1.31 (.09)
β	−0.11 (.12)	−0.04 (.13)
β^*	−0.24 (.01)	0.76 (.10)
Study 2 (w/judgment feedback)		
Low D		
σ	1.30 (.07)	1.30 (.05)
$\alpha = \sigma^*$	0.63 (.04)	0.76 (.04)
β	−0.23 (.08)	0.04 (.10)
β^*	−0.55 (.05)	1.19 (.06)
High D		
σ	1.17 (.06)	1.11 (.03)
$\alpha = \sigma^*$	1.18 (.06)	1.51 (.08)
β	−0.35 (.12)	−0.16 (.14)
β^*	−0.28 (.02)	0.60 (.03)

Note. BR, base rate; D, diagnosticity.

were chosen so that the predicted means matched the actual means (determining α and β), and the predicted pooled variance of the two conditional distributions matched the actual pooled variance (determining σ).

Table 2 lists the mean estimated parameter values in each condition, as well as the mean parameter values necessary for perfect calibration (σ^* and β^*). The most pronounced effect of the experimental manipulations is the influence of diagnosticity on α, which reflects the improved evidence quality in the high D conditions, and the extent to which participants were consequently able to use the cues to better discriminate stock price increases from decreases.

Recall that for perfect calibration, σ needs to exactly track α (i.e., $\sigma^* = \alpha$), and β must be responsive to the base rate (matching β^*). In Study 1, however, consistent with the view that judgment is primarily case-based and hence insensitive to these aggregate properties of the task environment, σ and β were only minimally sensitive to the manipulations of D and BR. The difference in observed β across the base rate manipulation is much smaller (−.20 vs. −.04) than the difference in β^* required to maintain good calibration (−.34 vs. 1.1; interaction is significant, $F(1,75) = 95.4$, $p < .0001$). In fact, the small effect of the base-rate manipulation on β failed to achieve statistical significance, $F(1,75) = 1.48$, $p > .20$. The difference in observed σ across the diagnosticity manipulation is in the wrong direction and trivially small (1.2 vs. 1.1, $F(1,75) = 2.69$, $p > .10$.) compared to the difference in σ^* required to maintain good calibration (.68 vs. 1.3; interaction is significant $F(1,75) = 54.2$, $p < .0001$).

The results of Study 2 (see bottom of Table 2) indicate that the relative insensitivity of β and σ remains even when feedback is presented following every probability judgment, indicating that additional outcome feedback by itself does not automatically improve calibration performance (see also, e.g., Baranski & Petrusic, 1994). The overall pattern of results in Study 2 is very similar to that of Study 1; the presence or absence of feedback did not interact with either the base rate or diagnosticity manipulations in influencing β or σ. How-

ever, several effects that did not achieve statistical significance in Study 1 did so in Study 2, possibly due to the larger sample sizes. The base-rate manipulation had a significant effect on β (-0.29 vs. -0.06), $F(1,90) = 4.46$, $p = .038$, although again the change in β was significantly smaller than the required difference in β^* (-0.42 vs. 0.90; interaction $F(1,90) = 84.8$, $p < .0001$) required for Bayesian judgment. Interestingly, the effect of the diagnosticity manipulation on σ was opposite to that required to maintain calibration (1.30 for low D and 1.14 for high D), $F(1,90) = 8.58$, $p < .01$. One possible interpretation is that the greater difficulty of the low D task led to greater noise or volatility in judges' responses. According to this interpretation, participants essentially chose to "play their hunches" more often, given the unpredictability of the domain.

The right-hand panel in Fig. 3 shows the calibration curves predicted by case-based RST and illustrates the sizeable miscalibration attributable to the failure of β and σ to match β^* and σ^*. Case-based RST with constant β and σ across conditions closely reproduces the general qualitative patterns of miscalibration observed in the four conditions. These predictions of the RST model require estimating a very small number of parameters; the parameters β and σ are estimated and held constant across all four conditions, and the discriminability parameter α is estimated within each level of diagnosticity.

Table 3 provides information on the quantitative fit of the RST model. The RST model was fit to each condition individually ("Free-Parameter RST"), and also fit to all the conditions subject to the constraints implied by case-based judgment, with fixed β and σ across all conditions ("Case-Based RST"). In each case, we measure the average absolute deviation between the actual calibration curves and the RST-predicted calibration curves, weighted by the actual response frequencies so that more reliable data points appropriately receive more weight. Also provided for comparison are the fit measures for each condition based on the optimal RST parameters, indicating ideal (Bayesian) calibration.

Table 3 indicates that RST provides a close absolute fit to the data, with overall average deviations of 3.8 percentage points, substantially superior to the predictions of ideal calibration (average absolute deviations of 11.8 percentage points). Second, note that there is only a very small improvement from fitting RST with "free" parameters (average deviation of 3.8) compared to the more restrictive case where the β and σ parameters are fixed (average deviation of 4.5). This pattern again supports the predictions of case-based judgment—that judgments are based primarily on the impression conveyed by the information seen as relevant to the specific case, with little incorporation of the relevant class-based information, such as overall evidence diagnosticity and base rate.

Discussion

Overall, case-based RST reproduced the observed patterns of calibration quite closely. Furthermore, the estimated individual-level parameter values across experimental conditions were quite consistent with a case-based model of probability judgment, in which aggregate class-based considerations receive little or no weight. Feedback following probability judgments did not change the qualitative pattern of results.

The observed patterns of calibration across the experimental conditions of Studies 1 and 2 suggest that people are neither consistently overconfident, nor consistently underconfident, nor consistently well-calibrated (see Fig. 3). Overestimation was found in the presence of a low outcome base rate, but underestimation was found in the presence of a high outcome base rate. Overextremity was pronounced in the presence of low evidence diagnosticity but not in the presence of high diagnosticity, where calibration was quite good. In contrast to the substantial variability in patterns of calibration across experimental conditions, there is relative stability in the RST parameters; to a first approximation, the values of the β and σ parameters remain essentially constant across substantial changes in outcome base rate and evidence diagnosticity. This is precisely the pattern expected if judgment is primarily case-based (see Table 1).

Studies 3 and 4

In Studies 1 and 2, case-based information (i.e., the cue values) changed from trial to trial, whereas the class-based factors were manipulated between-subjects. An alternative interpretation of the earlier results, then, is

Table 3
Fit measures for RST models for Studies 1 and 2, by condition

Condition	Free-parameter RST	Case-based RST	Perfect calibration
Low D, Low BR	2.15	3.27	14.55
High D, Low BR	3.84	4.18	6.78
Low D, High BR	5.38	5.76	15.27
High D, High BR	3.84	4.87	10.76

Note. Low BR, low base rate (40%); High BR, high base rate (70%); Low D, low diagnosticity ($\alpha = 0.62$); High D, high diagnosticity ($\alpha = 1.17$). Fit measure is average absolute Deviation, in percentage points, between predicted and actual calibration curves, weighted by response proportions.

that people's judgments may not be inherently case-based and hence insensitive to class-based factors, but rather that judgments tend to be insensitive to any unchanging task factor. In Studies 3 and 4, this possibility is tested by varying outcome base rate or evidence diagnosticity within the task presented to each participant.

Method

Participants

Participants were 165 undergraduate and MBA students at the University of British Columbia and the University of Florida. Randomly selected participants were paid based on the accuracy of their judgments, again based on the Brier score. Two participants' data were excluded from analysis because their responses were constant, and data for three more were excluded because their judgments correlated negatively with the outcome.

Design and procedure

In Study 3, participants completed a two-block experiment in which either D or BR varied between blocks. Each block, involving different types of companies, consisted of a training session followed by a set of judgment trials. Outcome feedback was provided after each of the 50 trials in the training session. This session was followed by 30 p(increase) judgment trials, without feedback. This process (training and prediction) was then repeated for a second set of companies of a different type, to signal to participants that the task conditions (specifically outcome base rate and cue diagnosticity) could differ from that of the first block.

The high BR, low D condition of Studies 1 and 2 served as a baseline condition encountered by all participants, and was labeled as a set of technology companies. For one group of participants ($n = 57$), the second condition (finance companies) was associated with higher cue diagnosticity, corresponding to the high BR, high D condition of Studies 1 and 2. For another group ($n = 58$), the second condition (retail companies) was associated with a lower base rate of stock price increases, corresponding to the low BR, low D condition of Studies 1 and 2.

In Study 4, the company type varied from trial to trial. That is, the two blocks of 50 training trials from Study 3 were mixed in a single 100-trial training session in Study 4. Similarly, the two blocks of 30 judgment trials from Study 3 were mixed in a single 60-trial judgment session in Study 4. One group of participants ($n = 20$) in Study 4 was presented with a mix of technology companies (high BR, low D) and finance companies (high BR, high D); another group ($n = 25$) was presented with a mix of technology companies (high BR, low D) and retail companies (low BR, low D). If noticeable trial by trial variation in outcome base rate or evidence diagnosticity is required for its use in judgment, then the design of Study 4 should reduce or eliminate the tendency to neglect class-based evidence. In all other respects, Studies 3 and 4 were identical in method and procedure.

Results

Calibration of probability judgments

The left-hand panel of Fig. 4 displays the group calibration curves for each condition of Studies 3 and 4.

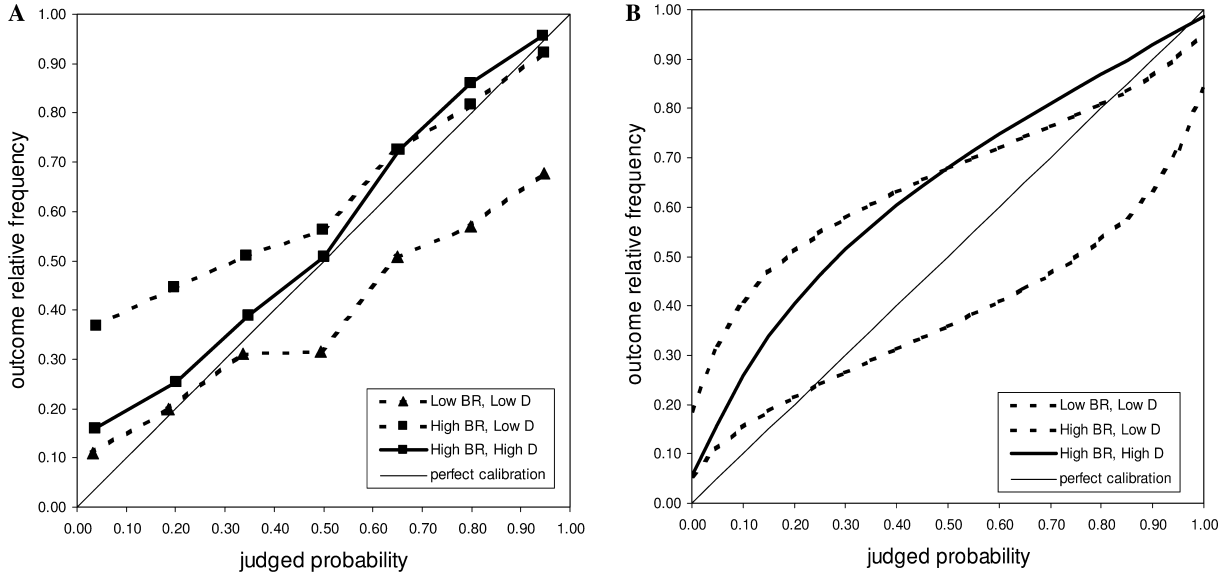

Fig. 4. Observed calibration curves (A) and curves predicted by case-based RST (B) for each combination of outcome base rate (BR) and evidence diagnosticity (D) in Studies 3 and 4. *Note.* Low BR = low base rate (40%); High BR = high base rate (70%); Low D = low diagnosticity ($\alpha = 0.68$); and High D = high diagnosticity ($\alpha = 1.09$). (B) RST parameter values β and σ are held constant across conditions, $\beta = 0.05$, $\sigma = 1.33$.

Data from the two studies have been collapsed into a single set of calibration curves; the general pattern is the same if the results of each study are plotted separately (comparisons of the average values of σ and β across studies yielded no significant differences, Fs < 3.0; see Table 4 also). As in Studies 1 and 2, the shape of the calibration curves in Fig. 4 is visibly influenced by both the base rate and diagnosticity manipulations. The base rate manipulation is again seen to influence the elevation of the calibration curves and the diagnosticity manipulation is seen to influence the slope.

RST model fit to data

The RST model was fit individually to each participant's data, using the method of moments procedure described earlier. Table 4 lists the mean estimated parameter values (σ, α, and β) in each condition of Studies 3 and 4, along with the mean parameter values necessary for perfect calibration (σ^* and β^*). Once again, consistent with the view that judgment is primarily case-based, σ and β were insufficiently sensitive to manipulations of D and BR, even when these class-based factors were manipulated within-subjects rather than between-subjects.

In Study 3, when base rate varied across blocks, the value of β was significantly higher in the high BR ($\beta = 0.37$) than in the low BR ($\beta = -0.11$) condition, $t(55) = 5.0$, $p < .001$. The observed values of β, however, are not nearly as large as those required to maintain good calibration, where $\beta^* = 1.65$ for high BR and $\beta^* = -0.56$ for low BR; the change in observed β is less than a quarter of that necessary to maintain good calibration, $t(55) = 8.6$, $p < .0001$. When diagnosticity varied across blocks, the observed value of σ did not change significantly, $t(52) = 1.17$, $p = .25$, and was too low relative to σ^* in the high D condition ($t(52) = 3.3$, $p < .01$) and too high in the low D condition ($t(52) = 4.9$, $p < .0001$). An unanticipated finding is that β is substantially higher in the high D condition than in the low D condition, $t(52) = 2.90$, $p < .01$. Based on the formula for β^*, it can be seen that the absolute value of β should actually *decrease* to maintain good calibration given an increase in α.

Table 4
Mean individual RST parameter estimate (and standard errors) for Studies 3 and 4, by diagnosticity and base rate conditions

Study 3 (blocked)	Low BR (w/Low D)	High BR (w/Low D)
σ	1.11 (.04)	1.09 (.05)
$\alpha = \sigma^*$	0.82 (.04)	0.64 (.04)
β	−0.11 (.10)	0.37 (.09)
β^*	−0.56 (.12)	1.65 (.16)
Study 3 (blocked)	Low D (w/High BR)	High D (w/High BR)
σ	1.05 (.05)	1.01 (.04)
$\alpha = \sigma^*$	0.81 (.04)	1.23 (.05)
β	0.21 (.11)	0.48 (.11)
β^*	1.17 (.12)	0.71 (.04)
Study 4 (mixed)	Low BR (w/Low D)	High BR (w/Low D)
σ	1.15 (.06)	1.18 (.05)
$\alpha = \sigma^*$	0.72 (.05)	0.90 (.06)
β	0.17 (.09)	0.14 (.10)
β^*	−1.29 (.13)	1.07 (.17)
Study 4 (mixed)	Low D (w/High BR)	High D (w/High BR)
σ	1.09 (.06)	0.95 (.06)
$\alpha = \sigma^*$	0.91 (.07)	1.38 (.10)
β	0.07 (.08)	0.32 (.12)
β^*	0.99 (.12)	0.78 (.12)

Note. BR, base rate; D, diagnosticity.

In Study 4, when base rate varied from trial to trial, the value of β did not differ significantly between the high BR ($\beta = 0.17$) and low BR conditions ($\beta = 0.14$), $t(27) = 0.29$, despite the pronounced difference in β^* between the low ($\beta^* = -1.29$) and high BR ($\beta^* = 1.07$); the difference $\beta - \beta^*$ changes significantly across conditions, $t(27)$=9.4, $p < .0001$. When diagnosticity varied from trial to trial, the observed value of σ changed significantly from low D ($\sigma = 1.09$) to high D ($\sigma = 0.95$) condition, but in the opposite direction of that required to maintain good calibration, $t(19) = 3.49$, $p = .002$.

The right-hand panel in Fig. 4 shows the calibration curves predicted by case-based RST (with β and σ constrained to constant values across both between and within-subject manipulations), which again closely reproduce the general patterns of miscalibration observed across experimental conditions. In terms of quantitative fit measures, the "free parameter" RST predictions were off by an average of 2.9 absolute percentage points; the predictions of case-based RST (where β and σ were held constant across conditions) were only slightly worse, off by an average of 3.7 percentage points. Both the constrained and unconstrained RST predictions were much better than the predictions of a perfectly calibrated Bayesian model, which had average absolute deviation of 9.7 percentage points. As in Studies 1 and 2, the case-based RST model captures the patterns in the data quite well; there is only very small improvement when the RST β and σ parameters are allowed to vary across different judgment conditions.

Study 5

The results of Studies 1–4 confirmed many of the predictions of case-based judgment, but one of the calibration patterns discussed earlier and illustrated in Fig. 1 has not been observed: underextremity. Case-based judgment predicts that underextremity will be found when diagnosticity is very high, and judges fail to adjust their probability judgments to incorporate this very high diagnosticity. Consequently, in Study 5, we manipulate diagnosticity over a wider range in order to test the

prediction of underextremity implied by case-based judgment.

Method

Participants

Participants were 161 undergraduate and MBA students at the University of British Columbia and the University of Florida. Eleven participants were dropped from the analysis because they did not use the cues appropriately, as evidenced by outlying negative ($n = 7$) or near-zero ($r < .15$, $n = 4$) correlations between judged probability and outcome.

Design and procedure

The overall structure of the experiment was similar to Study 2, with outcome feedback provided on all trials. There were three conditions, each with base rate of 50%, but differing in terms of cue diagnosticity. As in the earlier studies, in the low diagnosticity (low D) condition, the separation between increasing stock and decreasing stock cue distributions was 0.8 standard deviations (*SD*s) for domestic cues and 0.4 *SD*s for foreign cues. In the medium diagnosticity (medium D) condition, the separation was 1.2 *SD*s for domestic cues and 0.8 *SD*s for foreign cues (equivalent to the level of diagnosticity of the high D conditions in the previous studies). In the new high diagnosticity (high D) condition, the separation was 1.6 *SD*s for domestic cues and 1.2 *SD*s for foreign cues. If a simple performance measure is computed by adding total sales and subtracting total costs, the resulting value correlates 0.50, 0.68, and 0.82 with the dichotomous outcome variable, in the low D, medium D, and high D conditions, respectively.

Results

Calibration of probability judgments

The left-hand panel of Fig. 5 displays the group calibration curves for each level of diagnosticity. As before, the slope of the calibration curves is visibly influenced by the diagnosticity manipulation. Most importantly, note that the calibration curve shows overall underextremity in the high D condition. For each point along the high D calibration curve, subjective probabilities are less extreme than the corresponding objective probabilities.

RST model fit to data

The RST model was fit to each participant's data using the method of moments approach; Table 5 lists the mean estimated parameter values in each condition. Because the base-rate was 50% in all conditions, the optimal value of β^* is always 0.

The parameter σ ought to follow the parameter α for good calibration. However, the data show a roughly constant value of σ despite the changing level of α. In fact, σ decreases slightly as α increases, $F(2,147) = 3.41$, $p < .05$. The difference between σ and α changes significantly across conditions, $F(2,147) = 36.3$, $p < .001$, reflecting the failure to account for changes in diagnos-

Table 5
Mean individual RST parameter estimates (and standard errors) for Study 5, by diagnosticity condition (base rate constant at 50%)

	Low D	Med D	High D
σ	1.00 (.05)	0.94 (.05)	0.82 (.05)
$\alpha = \sigma^*$	0.84 (.06)	1.31 (.08)	2.29 (.17)
β	−0.61 (.11)	−0.72 (.20)	−0.55 (.13)

Note. D, diagnosticity.

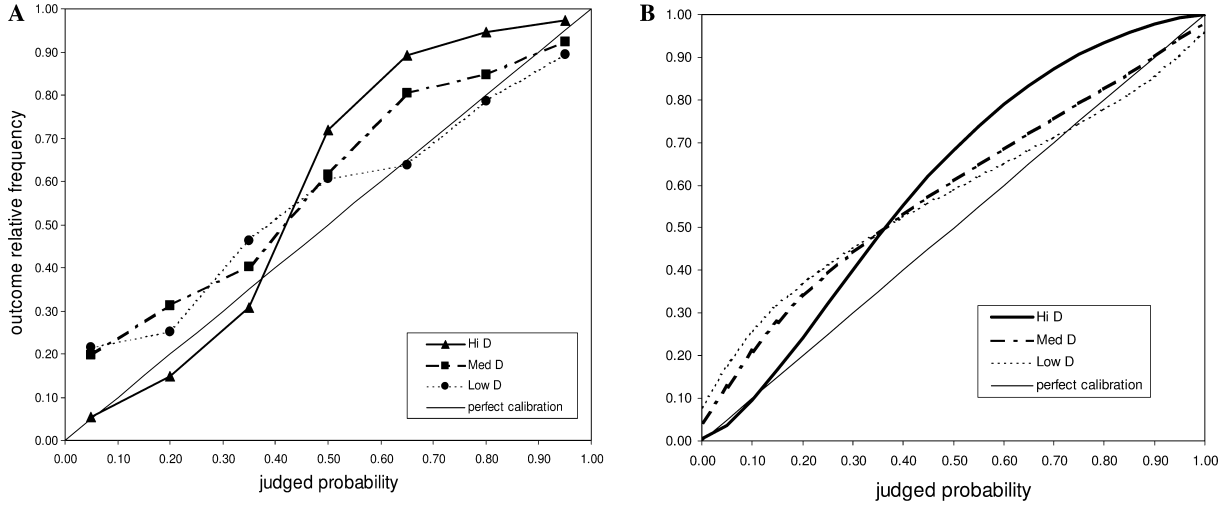

Fig. 5. Observed calibration curves (A) and curves predicted by case-based RST (B) for each diagnosticity (D) condition in Study 5. *Note.* Low D = low diagnosticity ($\alpha = 0.8$); Med D = medium diagnosticity ($\alpha = 1.0$); and High D = high diagnosticity ($\alpha = 1.7$). (B) RST parameter values β and σ are held constant across conditions, $\beta = -0.5$, $\sigma = 1.2$.

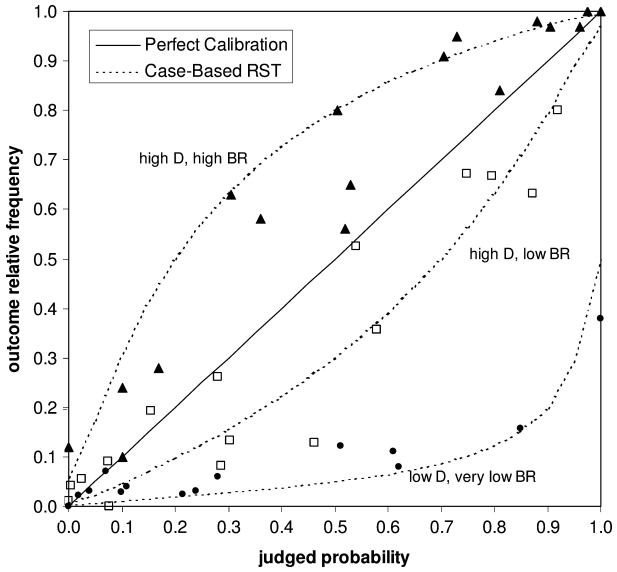

Fig. 6. Physicians. Calibration of physicians' judgments, varying by base rate and diagnosticity across studies. Circles represent very low BR and low D tasks, open squares represent low BR and high D tasks, and triangles represent high BR and high D tasks. Predictions of case-based RST model are displayed by dashed lines. See Koehler et al. (2002) for details. *Note.* Case-based RST predictions assume no focal bias ($\beta = 0$) and fixed judgmental extremity ($\sigma = 1$). Discriminability (D) and base rate (BR) are approximately matched to the empirical datasets as follows: For low D, $\alpha = 0.7$; for high D, $\alpha = 1.0$; very low BR = 5%; low BR = 30%; and high BR = 80%.

ticity. Consistent with the prediction of underextremity, σ is substantially too low for the high diagnosticity condition, $t(147) = 10.6$, $p < .001$.

The right-hand panel in Fig. 6 shows the calibration curves predicted by case-based RST and illustrates the sizeable miscalibration attributable to the failure of σ to follow α. In terms of the quantitative fit of the RST model, the free parameter RST predictions are off by an average of 4.1 percentage points, the case-based RST predictions are only slightly worse (off by 4.5 percentage points), and both are substantially better than the Bayesian predictions, which deviate by an average of 8.7 percentage points.

General discussion

Business students making probability judgments in a simulated stock market environment showed systematic patterns of miscalibration whether or not they received feedback after judgment trials and whether or not features of the environment remained constant or varied within the experimental session. Random Support Theory was able to closely reproduce these patterns, which are consistent with a case-based model of judgment developed in the heuristics and biases tradition (e.g., Gilovich, Griffin, & Kahneman, 2002; Kahneman, Slovic, & Tversky, 1982).

At first glance, these results and the accompanying analyses may seem like restatements of the extensively studied phenomena of "base-rate neglect" and "illusion of validity" (Kahneman et al., 1982). However, the present treatment provides substantial methodological, theoretical, and empirical advances over the classic heuristics and biases demonstrations.

First, it is again worth stressing that our results were obtained in a judgment environment where participants were able to learn cue diagnosticity and outcome base rate values directly from experience, and where judgmental accuracy was rewarded in an incentive-compatible manner. The demonstration of a clear, predictable pattern of biases in (a) an interactive setting with (b) an involving and easily understandable stock market task where participants (c) actively learn about environmental contingencies, and (d) have incentives for good performance helps to refute criticisms that results based on the scenario experiments used in classic heuristics and biases research have limited generality (e.g., Gigerenzer, 1991; Koehler, J., 1996; Schwarz, 1996). In particular, the observed pattern of results refutes the strong claim that biases of probability judgment "... are turning out to be experimental artifacts or misinterpretations." (Cosmides & Tobby, 1994, p. 327).

More specifically, a number of critics have argued that the results of scenario experiments in the heuristics and biases tradition do not accurately represent the quality of human statistical reasoning because the specific judgment task is unique and the relevant reference class is unclear. Some evolutionary psychologists have dismissed the results of such designs on the following logic: "our hunter-gatherer ancestors were awash in statistical information in the form of the encountered frequencies of real events: in contrast, the probability of a single event was inherently unobservable to them" (Cosmides & Tobby, 1994, p. 330). Thus it is notable that the pattern of miscalibration implied by case-based RST holds even when participants' beliefs are based on directly encountered frequencies, and their judgments are made repeatedly with salient and consequential feedback.

Second, the patterns in the observed calibration curves are inconsistent with models that propose that poor calibration is primarily a statistical artifact, ascribing miscalibration merely to the operation of random error or regression to the mean. Such models imply that shallow-sloping curves should cross the identity line at the base rate value (or at .50 if people are unaware of or unaffected by the base rate). Such models also cannot account for the underextremity found in Study 5.

Third, the results illustrate the usefulness and parsimony of RST as a quantitative model for predicting patterns of case-based judgment and as a tool for diagnosing observed deviations from perfect calibration. RST adds quantitative precision to the largely qualitative

principles that have emerged from the heuristics and biases tradition, which have sometimes been criticized for being too vague and imprecise (e.g., Gigerenzer, 1996). Furthermore, it extends the reach of support theory from the analysis of coherence of judgment to the analysis of correspondence between judgment and outcome.

Finally, and perhaps most importantly for our overall assessment of the performance of human probabilistic judgment, the observed patterns of calibration illustrate that overconfidence is not by any means a universal feature of probability judgment, but rather appears to be a common byproduct of case-based judgment—a byproduct likely to be observed only under some environmental circumstances. Note, however, that we are not arguing against the *existence* of optimistic overconfidence, confirmatory biases, or miscalibration resulting from random error. Indeed, we suspect that each of these processes may operate in some environments. These processes can also be diagnosed or captured by the parameters of RST (e.g., optimism may manifest itself as a larger value of β for desirable than for undesirable events). Most fundamentally, we argue that due to case-based judgment, evidential features of the environment will give rise to predictable patterns of (mis)calibration that cannot be explained easily by most other accounts.

Contrasting RST and similar stochastic models of calibration

The decision variable partition (DVP) model of Ferrell and McGoey (1980), supplemented with the assumption of fixed cutoff values, also can reproduce the different patterns of calibration across different environmental conditions. The primary modeling difference between RST and DVP is that DVP invokes multiple cutoff parameters (e.g., 10 cutoff parameters to model 11 response categories), whereas RST directly maps support into the observed quantitative judgment. In effect, the many DVP cutoff parameters are represented by the two RST parameters β (capturing the center of the cutoffs) and σ (capturing the spread of the cutoffs). As a result, RST has additional parsimony, but cannot be easily applied to non-quantitative judgments of likelihood, which DVP can model straightforwardly. With the additional parsimony of RST comes the useful result that there are unique values of β and σ corresponding to good calibration, whereas in models with multiple cutoffs, there are typically many sets of cutoffs that correspond to good calibration (see also Gu & Wallsten, 2001).

Perhaps the most fundamental difference between RST and DVP (and related signal-detection models with cutoff parameters), is that RST attaches the interpretation of *support* to the underlying random variable that is modeled. The principle of case-based judgment built from earlier findings in the heuristics and biases tradition furthermore implies that these assessments of support are primarily case-based. Thus, the stability of the β and σ parameters in RST has a core psychological interpretation in terms of case-based judgment. The stability of cutoff parameters in DVP, on the other hand, while equally valuable in terms of accounting for empirical changes in calibration performance, is less readily connected to a more fundamental principle of intuitive judgment. Hence, we suggest that the RST framework, with its linkage to qualitative psychological principles like case-based judgment, may provide a rich theoretical structure for representing calibration.

Generalization beyond the laboratory

Our results support the view that probability judgment is primarily case-based, resulting in neglect of class-based characteristics such as the general predictability of the environment and the overall outcome base rate. With the assumption that probability judgments are case-based, the RST model is able to capture large and systematic changes in calibration curves via *constant* values of the model's β and σ parameters (which can be compared to Bayesian parameter values β^* and σ^* needed for perfect calibration).

How readily might these results generalize beyond the laboratory setting? To address this question, Koehler et al. (2002) collected a number of previously published datasets consisting of on-the-job probability judgments made by experts in various domains such as medicine, meteorology, sports, economic forecasting, business, and law. Within each domain, the outcome variable of interest varied from dataset to dataset in terms of base rate and predictability. Consistent with our experimental findings, the elevation of the expert calibration curves varied systematically with outcome base rate, and the slope varied systematically with the diagnosticity of the available evidence. As an example, Fig. 6 displays the calibration of physicians' predictions across tasks differing in predictability and base rate, along with the case-based RST predictions of their performance. These patterns of miscalibration were quite closely reproduced by RST on the assumption that the extremity (σ) parameter remains constant across varying levels of evidence diagnosticity, and the β parameter remains constant across varying levels of outcome base rate (and diagnosticity as well).

Previous characterizations of subjective probability calibration, whether as consistently overconfident, underconfident, or well-calibrated, offer overly static portraits that fail to capture the predictable manner in which patterns of calibration change with properties of the judgment environment. Such characterizations could be modeled using RST on the assumption that its parameter values change across judgment environments in a manner that yields a fixed calibration pattern.

That is, fixed calibration patterns can arise from varying model parameters. A case-based judgment account, in contrast, can be modeled using RST on the assumption that the model parameters are largely insensitive to aggregate characteristics of different judgment environments. That is, fixed model parameters can yield varying patterns of calibration (illustrated in Figs. 1 and 6). Our data suggest that judgments both in the laboratory and in the field exhibit diverse patterns of calibration that are nevertheless largely predictable from the common fundamental principle of case-based judgment.

Appendix A. Determining Bayesian RST parameters

Recall the distributions of the logodds-transformed judged probabilities, conditional on each outcome, as described in Eqs. (4a) and (4b)

$$\ln\left(\frac{s(A_a)}{s(B_a)}\right) \text{ is normally distributed with mean } (\beta + \alpha)\sigma \text{ and variance } 2\sigma^2. \tag{4a}$$

$$\ln\left(\frac{s(A_b)}{s(B_b)}\right) \text{ is normally distributed with mean } (\beta - \alpha)\sigma \text{ and variance } 2\sigma^2. \tag{4b}$$

In this Appendix, we derive the values of $\beta = \beta^*$ and $\sigma = \sigma^*$ that will produce perfectly calibrated Bayesian judgments. Algebraically, these optimal values can be determined by matching the judged probability to the objective Bayesian probability of the outcome derived from the support distributions in ((4a) and (4b)). Consider an arbitrary judged probability $P(A,B) = \frac{1}{1+\exp(-j)}$ so that the judged log-odds are expressible as $j = \ln\left(\frac{P(A,B)}{1-P(A,B)}\right)$. We use Bayes's rule to determine the actual log-odds of hypothesis A being correct conditional on the judged log-odds j. As before, let the subscript a denote the event that hypothesis A is correct and the subscript b denote the event that hypothesis B is correct. Further, let $f_a(\cdot)$ and $f_b(\cdot)$ denote the density functions of the normal distributions of log-support specified in ((4a) and (4b)).

Bayes's rule in odds form implies:

$$\frac{\Pr(a|j)}{\Pr(b|j)} = \left(\frac{\mathrm{BR}}{1-\mathrm{BR}}\right) * \left(\frac{f_a(j)}{f_b(j)}\right). \tag{6}$$

Taking logs of both sides yields:

$$\ln\left(\frac{\Pr(a|j)}{\Pr(b|j)}\right) = \ln\left(\frac{\mathrm{BR}}{1-\mathrm{BR}}\right) + \ln\left(\frac{f_a(j)}{f_b(j)}\right). \tag{7}$$

Now, we plug in the specific normal distribution density functions $f_a(\cdot)$ and $f_b(\cdot)$, based on the means and standard deviations specified above. Note that the relevant standard deviation for the log-odds expression used is $\sqrt{2}\sigma$.

$$f_a(j) = \frac{1}{2\sigma\sqrt{\pi}} \exp\left(-\frac{1}{2}\left(\frac{j-\beta\sigma-\alpha\sigma}{\sqrt{2}\sigma}\right)^2\right), \tag{8}$$

$$f_b(j) = \frac{1}{2\sigma\sqrt{\pi}} \exp\left(-\frac{1}{2}\left(\frac{j-\beta\sigma+\alpha\sigma}{\sqrt{2}\sigma}\right)^2\right). \tag{9}$$

Substituting in the density functions:

$$\ln\left(\frac{\Pr(a|j)}{\Pr(b|j)}\right) = \ln\left(\frac{\mathrm{BR}}{1-\mathrm{BR}}\right) + \frac{1}{2}\left(\frac{j-\beta\sigma+\alpha\sigma}{\sqrt{2}\sigma}\right)^2 - \frac{1}{2}\left(\frac{j-\beta\sigma-\alpha\sigma}{\sqrt{2}\sigma}\right)^2. \tag{10}$$

Simplifying:

$$\ln\left(\frac{\Pr(a|j)}{\Pr(b|j)}\right) = \ln\left(\frac{\mathrm{BR}}{1-\mathrm{BR}}\right) + \frac{1}{4}\left(\frac{j}{\sigma}-\beta+\alpha\right)^2 - \frac{1}{4}\left(\frac{j}{\sigma}-\beta-\alpha\right)^2. \tag{11}$$

After expanding the two squared right-hand expressions and canceling duplicate terms:

$$\ln\left(\frac{\Pr(a|j)}{\Pr(b|j)}\right) = \ln\left(\frac{\mathrm{BR}}{1-\mathrm{BR}}\right) + \left(\frac{j}{\sigma}-\beta\right)\alpha. \tag{12}$$

We can then express the Bayesian log-odds, as a linear function of the judged log-odds j

$$\ln\left(\frac{\Pr(a|j)}{\Pr(b|j)}\right) = \left[\ln\left(\frac{\mathrm{BR}}{1-\mathrm{BR}}\right) - \alpha\beta\right] + \left(\frac{\alpha}{\sigma}\right)j. \tag{13}$$

For optimal calibration, this Bayesian log-odds expression needs to be equal to the judged log-odds j. We define β^* and σ^* as the parameter values for which this equality holds:

$$j = \left[\ln\left(\frac{\mathrm{BR}}{1-\mathrm{BR}}\right) - \alpha\beta^*\right] + \left(\frac{\alpha}{\sigma^*}\right)j. \tag{14}$$

For this equality to hold for all j, the bracketed intercept term must be zero,

$$\left[\ln\left(\frac{\mathrm{BR}}{1-\mathrm{BR}}\right) - \alpha\beta^*\right] = 0, \text{ and therefore } \beta^* = \frac{1}{\alpha}\ln\left(\frac{\mathrm{BR}}{1-\mathrm{BR}}\right). \tag{15}$$

Also, the slope term (the ratio α/σ^*) must be 1, and therefore $\sigma^* = \alpha$.

Thus, for perfect calibration the optimal extremity parameter σ^* must precisely follow (indeed, exactly equal) the discriminability parameter α, and the optimal bias parameter β^* must follow the outcome base rate (transformed to log-odds and scaled by α as well).

The constraint $\sigma^* = \alpha$ entails what may seem to be a surprising linkage between the degree of overlap between the means of the log-support distributions (α)

and the variability of those distributions (σ). The intuition for this constraint is that, as the separation between the two log-odds distributions α increases (or, equivalently, as the overlap between the two distributions gets smaller), the Bayesian probability $\Pr(a|j)$ becomes more extreme. To achieve good calibration then, one's judgment needs to get appropriately extreme as well; more extreme judgments entail a correspondingly larger value of σ. Consequently, Bayesian judgment in the RST framework requires that σ follow α. What is rather surprising, and results from the use of the log-normal distribution to represent support, is that σ must move in absolute lockstep with α.

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology, 35*, 303–313.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics, 55*, 412–428.

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

Brenner, L. A. (1995). A stochastic model of the calibration of subjective probabilities. Doctoral dissertation, Stanford University.

Brenner, L. A. (2000). Should observed overconfidence be dismissed as a statistical artifact? Critique of Erev, Wallsten, and Budescu (1994). *Psychological Review, 107*, 943–946.

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

Brenner, L. A., & Koehler, D. J. (1999). Subjective probability of disjunctive hypotheses: Local-weight models for decomposition of evidential support. *Cognitive Psychology, 38*, 16–47.

Brenner, L. A., & Rottenstreich, Y. (1999). Focus, repacking, and the judgment of grouped hypotheses. *Journal of Behavioral Decision Making, 12*, 141–148.

Brenner, L. A., Koehler, D. J., & Rottenstreich, Y. (2002). Remarks on support theory: Recent advances and future directions. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

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*, 173–188.

Cosmides, L., & Tooby, J. (1994). Better than rational: Evolutionary psychology and the invisible hand. *American Economic Review, 84*, 327–332.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*, 519–527.

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

Ferrell, W. R. (1994). Discrete subjective probabilities and decision analysis. In G. Wright & P. Ayton (Eds.), *Subjective probability*. Chichester: Wiley.

Fox, C. R. (1999). Strength of evidence, judged probability, and choice under uncertainty. *Cognitive Psychology, 38*, 167–189.

Fox, C. R., & Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management Science, 44*, 879–895.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond heuristics and biases. In W. Stroebe & M. Hewstone (Eds.). *European review of social psychology* (Vol. 2, pp. 83–115). Chichester, England: Wiley.

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

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review, 103*, 592–596.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*, 411–435.

Griffin, D., Gonzalez, R., & Varey, C. (2000). The heuristics and biases approach to judgment under uncertainty. In A. Tesser & N. Schwarz (Eds.), *Blackwell handbook of social psychology, Vol. 1: Intraindividual processes*. Oxford, UK: Blackwell.

Gu, H., & Wallsten, T. S. (2001). On setting response criteria for calibrated subjective probability estimates. *Journal of Mathematical Psychology, 45*, 551–563.

Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgment: The cyclical power model. *Psychological Review, 107*, 500–524.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57*, 226–246.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104*, 344–366.

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*, 384–396.

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

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

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

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*, 217–273.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes, 79*, 216–247.

Koehler, D. J. (1996). A strength model of probability judgments for tournaments. *Organizational Behavior and Human Decision Processes, 66*, 16–21.

Koehler, D. J. (2000). Probability judgment in three-category classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 28–52.

Koehler, D. J., Brenner, L. A., & 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*. 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.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Normative, descriptive and methodological challenges. *Behavioral and Brain Sciences, 19*, 1–53.

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

Liberman, V., & Tversky, A. (1993). On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin, 114*, 162–173.

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

Macchi, L., Osherson, D., & Krantz, D. H. (1999). A note on superadditive probability judgment. *Psychological Review, 106*, 210–214.

Massey, C. & Wu, G. (2005). Detecting regime shifts: The psychology of under- and overreaction. *Management Science*, forthcoming.

Novemsky, N., & Kronzon, S. (1999). How are base-rates used, when they are used: A comparison of additive and Bayesian models of base-rate use. *Journal of Behavioral Decision Making, 12*, 55–69.

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

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104*, 406–415.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*. Mahwah, NJ: Erlbaum.

Shiller, R. J. (2000). *Irrational exuberance*. Princeton, NJ: Princeton University Press.

Sloman, S. A., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 573–582.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes, 65*, 117–137.

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

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*, 547–567.

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

Wallsten, T. S., & Budescu, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science, 29*, 151–173.

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

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39*, 806–820.

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