ACADEMIC PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Organizational Behavior and Human Decision Processes 90 (2003) 87–110

ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES

www.elsevier.com/locate/obhdp

A random support model of the calibration of subjective probabilities

Lyle A. Brenner

Department of Marketing, Warrington College of Business Administration, University of Florida, Gainesville, FL 32611-7155, USA

Abstract

A stochastic model of the calibration of subjective probabilities based on support theory (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994) is presented. This model extends support theory—a general representation of probability judgment—to the domain of calibration, the analysis of the correspondence between subjective and objective probability. The random support model can account for the common finding of overconfidence, and also predicts the form of the relationship between overconfidence and item difficulty (the "hard–easy effect"). The parameters of the model have natural psychological interpretations, such as *discriminability* between correct and incorrect hypotheses, and *extremity* of judgment. The random support model can be distinguished from other stochastic models of calibration by: (a) using fewer parameters, (b) eliminating the use of variable cutoffs by mapping underlying support directly into judged probability, (c) allowing validation of model parameters with independent assessments of support, and (d) applying to a wide variety of tasks by framing probability judgment in the integrative context of support theory.

1. Introduction

The study of subjective probability, from both normative and descriptive perspectives, has long concerned psychologists, statisticians, economists, and other behavioral and social scientists. People use subjective probabilities to represent their beliefs about the likelihood of future events or their degree of confidence in the truth of uncertain propositions. Consequently, understanding the nature of subjective probability allows a glimpse into the structure of human knowledge and belief. Furthermore, because the assessment of likelihood is an essential component of choices made under uncertainty, developments in the study of subjective probability are applicable to many models of decision-making.

One facet of the descriptive study of judgment under uncertainty concerns the *calibration of* subjective probabilities: how well do subjective probabilities match corresponding objective probabilities? Quite apart from the psychological importance of subjective probability as a measure of belief, the practical question of the

E-mail address: lbrenner@ufl.edu.

0749-5978/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0749-5978(03)00004-9

relationship between subjective and objective probability is of great importance to many applied endeavors, particularly risk and decision analysis.

1.1. Calibration studies

In a typical psychological study of calibration, a participant answers a number of questions, or makes a series of forecasts about future events, and for each item expresses a subjective probability that the chosen answer or forecast is correct. A person is considered well-calibrated if for all events assigned a given subjective probability p, $100p\%$ of the events occur as predicted. Ideal calibration entails a precise match between subjective assessments of likelihood and the corresponding empirical relative frequencies. The psychological literature on calibration is vast. Several quite comprehensive discussions of empirical findings and analytic methods in the study of calibration are provided by Keren (1991), Lichtenstein, Fischhoff, and Phillips (1982), McClelland and Bolger (1994), Spetzler and Staël von Holstein (1975), Wallsten and Budescu (1983), Wallsten (1996), and Yates (1994).

1.2. Overconfidence and the hard–easy effect

A common, though by no means universal, finding from many studies of calibration is that people are often *overconfident*; subjective probabilities are frequently more extreme than corresponding accuracy rates. For example, when people express 95% confidence, they may be correct only about 80% of the time.

Another common empirical pattern has been termed the *hard–easy effect* (also known as the difficulty effect): the degree of overconfidence is larger for difficult tasks than for easy tasks. Difficulty of a task can be measured either by the proportion of judges identifying the correct answer, or by subjective assessments of difficulty made by the judges or by others. There has been substantial debate about the interpretation and generality of both overconfidence and the hard–easy effect, with some arguing that these patterns may, in part, represent spurious effects or statistical artifacts (e.g., Ariely et al., 2000; Erev, Wallsten, & Budescu, 1994; Gigerenzer, Hoffrage, & Kleinblölting, 1991; Juslin, 1994; Juslin, Winman, & Olsson, 2000; Klayman, Soll, Gonzalez-Vallejo, & Barlas, 1999; Soll, 1996).

In this paper, I introduce the random support model of calibration, a stochastic model of calibration based on support theory (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994). The random support model predicts the hard–easy effect, and can account for the frequent finding of overconfidence; it can also account for the less common finding of underconfidence. Although the random support model cannot itself resolve the various debates concerning the interpretation and generality of these various empirical patterns, the model can nonetheless account for and offer a parsimonious representation of them. Furthermore, the parameters of the model permit psychologically meaningful comparisons between the properties of probability judgments across different experimental manipulations, task types, or judge characteristics.

The random support model shares several features in common with other stochastic models of probability judgment. Like the decision variable partition model of Ferrell and McGoey (1980), the random support model treats subjective certainty as a random variable, and incorporates different distributions for subjective certainty in true and false propositions. Like the stochastic judgment model of Wallsten and Gonzalez-Vallejo (1994), the random support model explicitly incorporates variability (often termed "error") in the judgment process, as do models proposed by Erev et al. (1994), Soll (1996), Pfeifer (1994), and Dougherty, Gettys, and Ogden (1999). Detailed comparisons between the random support model and several other similar models can be found in the final section of the paper.

1.3. Overview

The paper is organized as follows. I first review support theory, on which the random support model is based, introduce the stochastic model of probability judgment, and apply the model to two-alternative forced choice calibration tasks. The model makes specific quantitative predictions about the (nonlinear) relation between overconfidence and item difficulty that is observed in the data. In addition, independent assessments of support for particular hypotheses are closely related to parameter estimates from the fit of the random support model, further validating the approach.

Next, the interpretations of the model parameters are illustrated by fitting the model to several existing data sets, and comparing the estimated parameter values across different conditions.

The model is then applied to three-alternative forced choice tasks. Data from pairwise comparisons of US state populations from one group of judges are used to generate model predictions for three-alternative judgments from a different group of judges. The success of these predictions demonstrates the model's ability to account for independent data from different tasks.

Finally, the random support model is compared to other models of calibration, including previous stochastic models. Theoretical and methodological implications for the interpretation of patterns in calibration data, and for interpretations of variability in judgment, are discussed.

2. Random support model for multi-alternative tasks

The random support model is an extension of support theory (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994), a general model of subjective probability. Support theory is premised on the critical distinction between events (as subsets of some sample space) and descriptions of those events, which are termed *hypotheses*. In support theory, judgments of probability are attached to hypotheses rather than events; as a consequence, different descriptions of the same event are allowed to yield different judged probabilities. Let $P(A, B)$ denote the judgment of the probability that hypothesis A holds, rather than hypothesis B, assuming A and B describe disjoint events exactly one of which obtains. The first argument A is termed the focal hypothesis, and the second argument B the alternative hypothesis. Support theory assumes that there exists a support scale $s(\cdot)$ such that probability judgment can be represented as normalized support:

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

The support for a hypothesis can be interpreted as a measure of the overall strength of the evidence for the hypothesis. This degree of support may be based on a variety of psychological processes, including retrieval of past empirical frequencies, judgmental heuristics such as, availability or representativeness, or logical/mathematical arguments or deductions (cf. Brenner & Koehler, 1999; Brenner, Koehler, & Rottenstreich, 2002; Koehler, Brenner, & Tversky, 1997; Rottenstreich, Brenner, & Sood, 1999). Past work in support theory has focused in large part on how the specificity of a description may affect its judged likelihood; in particular, unpacking an aggregate hypothesis (e.g., "unnatural cause of death") into a collection of components (e.g., "homicide, suicide, accident, or some other unnatural cause of death") tends to increase support for the hypothesis. Although the random support model is able to address such effects in a relatively straightforward way (by introducing different distributions of support for packed and unpacked hypotheses), the

present goal is to apply the support theory representation given in Eq. (1) to the modeling of calibration. This allows the theory to be applicable to any set of hypotheses for which the outcomes are observable, not only for hypotheses reflecting alternative descriptions of the same event.

2.1. Random support model of probability judgment

Support theory as currently described is a deterministic model, in which the support value for a given hypothesis is represented as a constant. However, support for a particular hypothesis can be naturally thought of as subject to variability from several sources. For example, different judges evaluating a particular hypothesis are likely to draw on different sets of evidence (between-judge variability of support). Furthermore, the same judge evaluating a hypothesis at different times or under different circumstances is also likely to draw on different sets of evidence (within-judge variability of support).

To accommodate variability of support, the present model of calibration treats support as a random variable, and specifies probability distributions of support for correct and incorrect hypotheses. The predicted calibration function can then be derived based on these distributions. The overall approach is similar in spirit to the decision variable partition model of Ferrell and McGoey (1980), which proposed distributions of "subjective certainty" for correct and incorrect responses, and derived the calibration function from these distributions along with a set of cutoffs for converting subjective certainty into a probability judgment. A detailed comparison of the random support model with the decision variable partition model is presented in the general discussion.

2.1.1. Preliminaries

Before describing the central aspects of the random support model, it will be helpful to define some terms and notation. First, consider transforming judged probability (e.g., 75%) into "odds" (e.g., 3–1). Given the support theory representation, the odds of hypothesis A rather than B is easily shown to be the ratio of the hypotheses' support values:

$$R(A,B) = \frac{P(A,B)}{1-P(A,B)} = \frac{s(A)}{s(B)}. \tag{2}$$

Also define the log-odds of A rather than B:

$$L(A,B) = \ln R(A,B) = \ln s(A) - \ln s(B). \tag{3}$$

2.2. Applying the random support model to calibration

Consider a sample question to be evaluated for calibration quality: "Which state has more inhabitants: Arizona or Wisconsin?" In a two-alternative forced choice (2AFC) task like this, the subject chooses an answer and then assesses the probability that the chosen answer is correct.

In answering such a question, each hypothesis ("Wisconsin more populous than Arizona" and "Arizona more populous then Wisconsin") is assumed to receive some degree of support, based on the subject's knowledge of the question domain. Furthermore, support for each hypothesis is assumed to vary across different judges, as well as within a particular judge considering a particular hypothesis at different times. For example, to judge a state's population, Adam may think of familiar cities that are in that state. On one occasion, he may think of Milwaukee and Green Bay as Wisconsin cities; on another occasion, Madison, Eau Claire, and Sheboygan may also be recalled in addition to the other two cities, and Adam's support for a large

population in Wisconsin would increase. Different evidence favoring particular cues may come to mind on different occasions, as in the example above, or entirely different predictive cues may be used in the determination of support.

In addition to variability of evidence considered by a particular judge, the type and quantity of evidence considered by different judges will also vary. Eve may have very different evidence than Adam about Wisconsin cities, or may use different cues to assess state populations (e.g., considering number of professional sports teams or number of acquaintances from Wisconsin rather than the number of remembered cities). The probability distributions of support that underlie the random support model may be used to describe both within-judge and between-judge variability of support. It is up to the researcher to decide over what set of judgment occasions to apply the model. Note that this approach allows the model to be very flexible in its application, and makes no rigid assumptions about the nature of variability of support. This approach does require, however, that in interpreting results from the model, researchers are careful in specifying what type(s) of variability they are attempting to account for.

The random support model of calibration has two primary components: a rule for converting support to judged probability, provided by support theory Eq. (1), and a joint probability distribution of support for the correct and incorrect answers. Together, these two components allow the model to predict how often a judge will choose the correct hypothesis, and also to predict the distribution of probability judgments in the interval $[.5, 1]$ they will assign to the chosen hypothesis. I consider the model's predictions of the judge's choice of answer and probability judgment in turn.

Choice of answer. The judge is assumed to choose the hypothesis which has greater support in a particular exposure to the problem. Given that support for each hypothesis is characterized by a random variable, this leads to a specific case of Thurstone's (1927) law of comparative judgment; in this particular case, two hypotheses are compared in terms of degree of support. Using the distributions of support for the hypotheses considered, the probability of choosing a particular hypothesis can be determined.

Let C denote the correct hypothesis, and I denote the incorrect hypothesis for a particular question. Let the random variable X denote the support for the correct answer; i.e., $X = s(C)$. For the Arizona–Wisconsin example, X represents the support for the (true) proposition "Wisconsin has more inhabitants than Arizona." Similarly, let the random variable $Y = s(I)$ denote the support for the incorrect answer. The researcher is assumed to know which hypothesis corresponds to C and which to I, although the judge of course does not.

When $X > Y$ the judge will choose the correct answer; when $X < Y$, the judge will choose the incorrect answer. (Due to the use of continuous support distributions, the probability of a "tie" between the support values is 0.) Given a particular joint distribution of support, the probabilities of these two events can be expressed in terms of a parameter representing the judge's ability to discriminate between correct and incorrect answers, as shown below.

It is important to distinguish between the *choice probabilities*, e.g., $Pr(C$ chosen over $I)$, and the *judged probabilities*, e.g., $P(C, I)$. I will use $Pr(\cdot)$ to denote choice probability, and $P(\cdot, \cdot)$ to denote judged probability. The former represents the probability of a judge selecting one option as the more probable outcome, and can be measured across judges. The latter represents an individual judge's subjective assessment of the probability that the chosen hypothesis is correct.

Probability judgment. Once the choice of answer is made, the judge assesses the probability that the chosen answer is correct. Because the judge chooses the option with the greater support, the judged probability that the chosen answer is correct is the normalized *maximum* of the two support values: $\max(X, Y)/(X + Y)$.

2.3. Predicting the calibration curve

Based on the distributions of support for the correct and incorrect answers, the probability of choosing the correct answer can be determined, as can the probability distribution of the judged probability. Furthermore, the distributions of judged probability conditional on the choice of the correct or incorrect hypothesis can also be determined. Using Bayes's rule, one can then derive the objective probability of making a correct choice given a probability judgment in a particular range, thereby specifying the calibration curve. The probability of choosing the correct hypothesis given a probability judgment between j and k is given by:

$$P(C \mid j < P < k) = \frac{\Pr(j < P < k \mid C)\Pr(C)}{\Pr(j < P < k)} = \frac{F_{P \mid C}(k) - F_{P \mid C}(j)}{F_P(k) - F_P(j)} \Pr(C) \tag{4}$$

In this equation, F_P denotes the cumulative distribution function (cdf) of P and $F_{P \mid C}$ denotes the conditional cdf of P given the choice of the correct answer. Both of these functions are derived from the initial distributions of support for correct and incorrect hypotheses.

2.4. Summary of model

The random support model of calibration consists of two main components: (a) a specification of the distributions of support corresponding to correct and incorrect hypotheses, and (b) a rule for converting support values to probability judgments.

This structure of the model makes clear a useful distinction between a model of probability *judgment* and a model of probability *calibration*. The support theory representation provides the model of probability judgment, which can be applied very generally, whether one is concerned with calibration or not.

The support distributions for correct and incorrect hypotheses allow the model to predict the calibration of probability judgments. These distributions instantiate the knowledge of the judges, and allow a link between a judge's statement of probability and the empirical likelihood that the chosen answer is correct.

2.5. Lognormal distributions of support

I will use the *lognormal* distribution to model variability of support throughout the remainder of the paper. The lognormal distribution is a close cousin of the familiar normal (or Gaussian) distribution. If U is a random variable following the normal distribution with mean μ, and variance σ^2, then $V = e^U$ follows a lognormal distribution with parameters μ and σ^2. For convenience, we use the same parameters μ and σ^2 to characterize the lognormal random variable, although these parameters no longer represent the mean and variance of V.

Using the lognormal distribution for support is attractive for several reasons. First, the lognormal distribution is positively skewed, with a long right-hand tail, which appears to be appropriate for many of the tasks to be considered. This shape implies that for a particular distribution of support values, most are relatively small, with a small proportion being quite large. In modeling variability across judges, this shape would capture the situation in which a few judges have a high degree of (perceived) knowledge, while many judges have a small or moderate amount of (perceived) knowledge.

A second reason for using the lognormal distribution is that, because of its close relationship to the familiar normal distribution, the lognormal distribution is

mathematically very convenient. Lognormal distributions of support lead to normal distributions of log-odds, and as a result a simple transformation from judged probabilities to log-odds will produce normally distributed data. Finally, and most importantly, using lognormal distributions of support provides a good fit to a variety of data. I wish to make no general claims regarding the superiority of the lognormal distribution across all circumstances and contexts. Other distributional families, such as the exponential and beta distributions, also have shown quite good fit to a variety of data. However, based on its success in fitting calibration data, and its close relationship to the familiar normal distribution, the lognormal distribution is a natural choice for modeling variability of support.

Independent, equal-variance log-support distributions. Let C and I denote correct and incorrect hypotheses, respectively. Consider $X = s(C)$ and $Y = s(I)$ to be independent lognormal random variables, summarized mathematically as $X \sim \Lambda(\mu_x, \sigma^2)$ and $Y \sim \Lambda(\mu_y, \sigma^2)$. Equivalently, $\ln\ X$ and $\ln\ Y$ can be characterized as independent normal random variables, with means μ_x and μ_y, respectively, and common variance σ^2. The interpretations in terms of log-support will typically be more natural for most readers, given the familiarity of the normal distribution.

Using familiar properties of independent random variables, the log-odds term $L(C,I)$ follows the normal distribution with mean $\mu_x - \mu_y$ and variance $2\sigma^2$. Equivalently, the odds expression $R(C,I)$ follows a lognormal distribution, $R(C,I) \sim \Lambda(\mu_x - \mu_y, 2\sigma^2)$. The judged probability $P(C,I)$ does not follow any familiar distributional family; however, its cumulative distribution is easily determined by transforming to log-odds and using the distribution of $L(C,I)$.

Correlated support distributions. In many cases, it may be appropriate to consider the support for the focal and the alternative hypotheses to be correlated. A negative correlation would most often make sense (e.g., if the two support assessments are based on the same cues). For example, if C is the hypothesis that "Wisconsin has a larger population than Arizona" and I is the hypothesis that "Arizona has a larger population than Wisconsin," then when the support for C is high, the support for I is likely to be low.

To model the case of correlated support, consider $\ln\ X$ and $\ln\ Y$ to follow a multivariate normal distribution with mean vector

$$\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$$

and covariance matrix

$$\begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}.$$

In this case,

$$L(C,I) \sim N\big(\mu_x - \mu_y, 2\sigma^2(1-\rho)\big).$$

Because support is an unobservable quantity (in essence derived from probability judgments, much as utility is derived from choices), there is no way to separate the parameters ρ and σ^2. For any $\rho < 1$, an empirically equivalent representation using independent support distributions (with appropriate variance) can be created. Thus, I will consider the simpler case of independent support distributions throughout the paper. This is a limiting assumption only when attempting to relate support derived from probability judgments to independent assessments of support. In addition, the successes of past investigations which have related separate assessments of hypothesis strength to derived support suggest that independent support distributions may often be a reasonable assumption (Fox, 1999; Koehler, 1996).

Now consider a more useful specification of these distributions, in which the two primary parameters that characterize the model are introduced. Let X and Y be independent lognormal random variables representing the support for the correct and incorrect answers for a particular question. Define the distributions of X and Y as follows: $X \sim \Lambda(\delta\sigma, \sigma^2)$ and $Y \sim \Lambda(0, \sigma^2)$. Equivalently, $\ln X \sim N(\delta\sigma, \sigma^2)$ and $\ln Y \sim N(0, \sigma^2)$. In this formulation, log-support for the correct and the incorrect answers follow equal-variance normal distributions, with the mean of the correct answer distribution shifted by δ standard deviations from the mean of the incorrect answer distribution.

The parameter δ thus represents the discriminability between the two distributions: the additional support that typically applies to correct hypotheses rather than incorrect hypotheses. Using properties of the normal distribution, we find that the probability of choosing the correct answer over the incorrect answer is a monotonic function of δ:

$$Pr(C) = Pr(X > Y) = Pr(\ln X > \ln Y) = Pr(\ln X - \ln Y > 0) = \Phi\left(\frac{\delta\sigma}{\sigma\sqrt{2}}\right) = \Phi\left(\frac{\delta}{\sqrt{2}}\right), \tag{5}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

The second parameter σ represents the degree of variability of the support distributions—the more spread out the support values, the greater the value of σ. In terms of the observable probability judgments rather than the unobservable construct of support, σ reflects the *extremity* of judgments; greater values result in judgments closer to 0 and 1, and further away from .5.

In the case of 2AFC tasks where judgments cannot be less than .5, larger σ entails larger probability judgments. Put another way, when σ is large, there is greater variability in the support distributions, and consequently, highly discrepant support values for the two considered options are more common, resulting in larger probability judgments for the selected option.

Examples. Fig. 1 displays the model's predictions for different values of δ and σ. The top curves represent calibration curves, and the bottom curves display the "response proportions" or relative frequencies of the various possible probability judgments. Note that the Y-coordinates of each of the response proportion curves sum to 1.

Larger values of δ result in an upward shift of the calibration curves (top portion of Fig. 1) and relatively more high probability judgments (bottom portion of Fig. 1), as can be seen by comparing the filled-circle and open-circle curves. Larger values of σ cause an overall drop in the calibration curve, and a greater proportion of high probability judgments, as can be seen by comparing the cross and the open-circle curves.

Even though changes in σ do not affect the overall likelihood of choosing the correct answer (because that likelihood depends only on δ—see Eq. (5)), changing σ can nonetheless greatly affect the height and shape of the calibration curve. In particular, changing the overall distribution of judgments while holding overall accuracy constant will affect the height of the calibration curve. As an illustration, consider a case where two points on calibration curve A are $(.60, .60)$ and $(.70, .70)$. Imagine that overall accuracy remains constant, but σ is increased such that probability judgments that previously were .60 become judgments of .70, and judgments that previously were .70 become judgments of .80. As a result, two new points on calibration curve B will be $(.70, .60)$ and $(.80, .70)$. The result is that curve B will be below curve A even though overall accuracy is unchanged. Thus, the calibration curve is dependent on both the discriminability between correct and incorrect answers (δ), and the variability of support (σ).

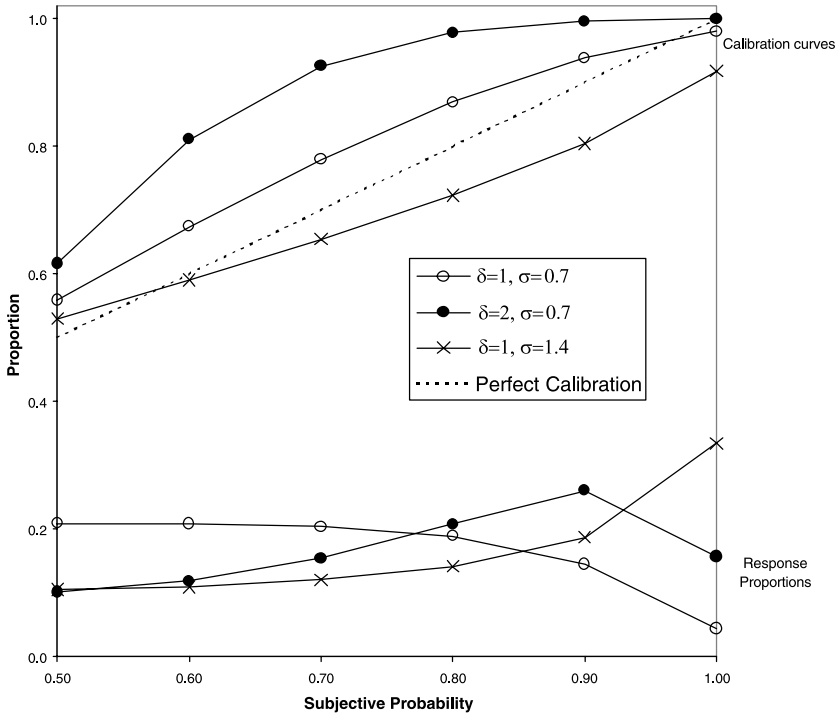

Fig. 1. Examples of calibration curves and response proportions based on lognormal-support model.

Parameter estimation. There are several ways in which the parameters of the random support model could be estimated from calibration data. A simple and effective one is a "method of moments" approach, which works as follows: δ is estimated from the proportion of correct choices made, using Eq. (5); σ is then estimated by equating the average probability judgment in the data with the average probability judgment implied by the model. Simple analytic expressions for the average judged probability cannot be determined for the lognormal model, so an iterative procedure for estimating σ was used. Because of its intuitive simplicity and computational ease, this estimation method will be used throughout the paper. Other approaches (method of moments in log-odds metric, maximum likelihood) give very similar results.

3. Study 1: Pairwise state population comparisons

To illustrate various analyses possible with the random support approach, the model was fit to data involving comparative judgments of the populations of states in the US.

3.1. Method

Twenty five pairs of states were randomly selected from the set of all possible pairs of states. Participants ($n = 190$ Stanford University undergraduates) evaluated these 25 state pairs (e.g., Kentucky and Utah), for each pair selecting the state they believed had the larger population. After selecting a state, each participant rated the probability that the choice was correct by circling a value on a scale ranging from 50% to 100% by 5% increments.

Item level and aggregate level fitting. The random support model can be fit at either an aggregate or an item level. In the case of aggregate fitting, two parameters

are estimated to match the overall accuracy rate and average probability judgment, across all state pairs and all judges. In other words, two parameters are fit to the entire set of data. In the item-level case, the model is fit to data from each individual judgment item (pair of states), across judges. In other words, two parameters are fit for each item. The model predictions across the 25 items can then be combined, and compared against the overall (across-item) data. It is of course also possible to fit the model individually to each judge's data, but with only 25 judgment items per judge, the calibration curves are rather "noisy" for each individual judge.

These various fitting procedures differ in terms of the sources of variability represented by the distributions of support. In the aggregate analysis, the distributions of support are used to account for within-judge, between-judge, and between-item variability. In the item-level analysis, the distributions of support account for only within-judge and between-judge variability, whereas between-item variability is captured by different fitted parameter values for each item.

The random support model fits approximately equally well to the full set of data via both the item-level and the aggregate fitting procedures. Fig. 2 displays the data (both calibration curve and response proportions, plotted as open triangles) and the predictions based on the estimated aggregate model (plotted as filled circles). The corresponding item-level graph is virtually identical.

The model's predictions closely approximate both the observed calibration curve and the response proportions. In terms of goodness-of-fit measures, the model is off by about 3 percentage points in predicting the calibration curve; specifically, the average absolute deviation between the predicted and observed calibration curves (weighted by the response proportions, to avoid overemphasizing differences based on little data) is .031 for the aggregate model, and .030 for the item-level model. As for the response proportions, the average absolute deviation between the predicted and observed proportions is .024 for the aggregate model, and .017 for the item-level

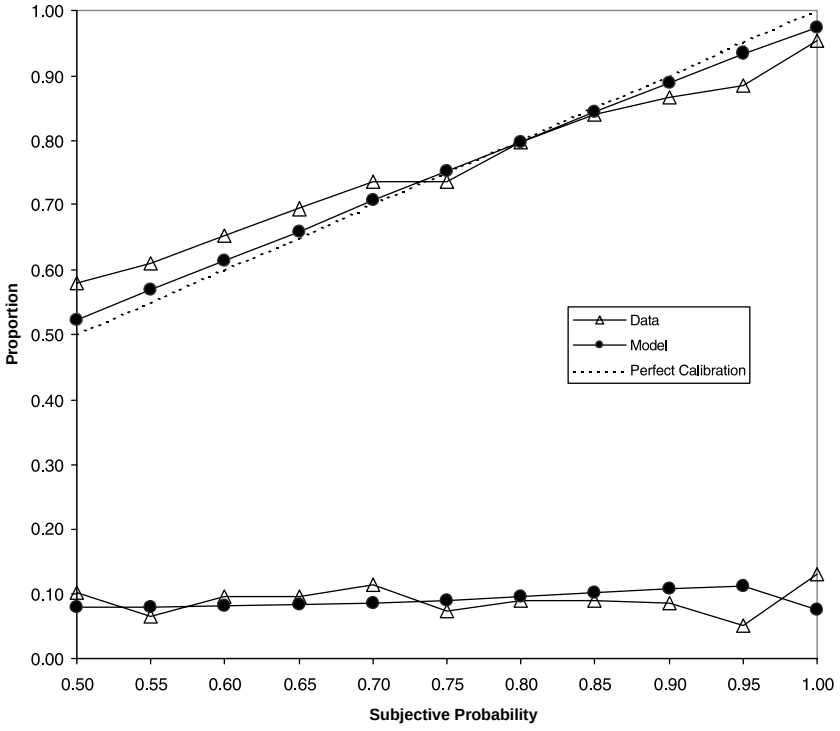

Fig. 2. Calibration curves and response proportions for aggregate data (triangles) and model predictions (circles) for Study 1 state population judgments.

model. The item-level model fits the response proportions a bit better, mainly because the predicted proportion of 100% judgments is increased. The two models make essentially identical predictions about the calibration curves. Descriptive measures of goodness-of-fit (like average absolute deviation) are desirable here because they can indicate the overall quality of fit without specifying any particular (arbitrary) null hypothesis to be tested. Furthermore, such measures are not affected by design features like sample sizes as hypothesis testing measures would be, nor are they contaminated by range restriction as correlational measures would be.

The similarity of the item-level and aggregate fits does not imply that there are no systematic differences in choices or probability judgments across the judgment items. Rather, it suggests that across-item variability can be incorporated effectively into the distributions of support that initially were used to account for variability of support across judges.

Item-level analyses. I now turn to more detailed analyses of the parameter estimates for individual judgment items. In particular, can differences in performance and judged confidence on different items be accounted for parsimoniously by the random support model? Can the model accommodate the hard–easy effect, the commonly found relationship between difficulty and overconfidence for individual items?

In Fig. 3, overall accuracy (i.e., the proportion of judges choosing the correct answer) is plotted against average judged probability for each of the 25 state pairs. The smooth curve in the figure is based on the lognormal random support model with constant $\sigma = .80$ as discriminability δ is varied. The correspondence between the curve and the distribution of points suggests a good qualitative fit of the lognormal model with constant σ across items. That is, average judged probability and overall accuracy in individual items are quite predictable based on only changes in

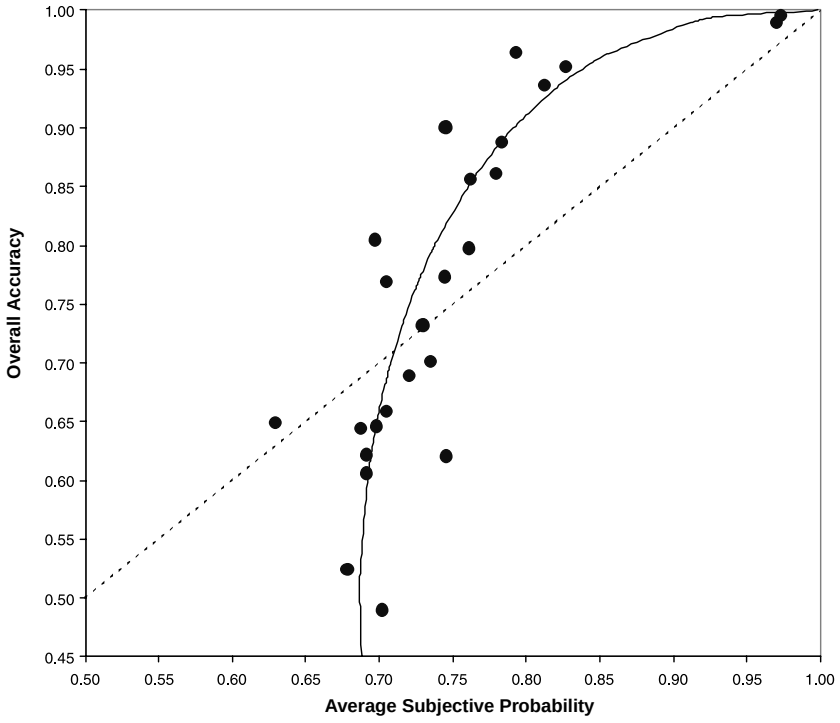

Fig. 3. Overall accuracy plotted against mean subjective probability for each pair of states in Study 1. The solid line traces the model predictions for varying discriminability δ and constant $\sigma = .80$.

the discriminability parameter δ, representing the difficulty of the individual questions.

Relation between overconfidence and difficulty. Holding constant the value of σ, the model can predict the relationship between accuracy and overconfidence. To illustrate the random support model's predictions of the hard–easy effect, overconfidence is plotted against accuracy for several different values of σ in Fig. 4. The random support model accounts for the hard–easy effect for difficult and moderately easy questions; however, it predicts an intriguing reversal of the hard-easy effect for extremely easy questions (accuracy of .9 or higher). In the context of the lognormal-support model, overconfidence drops as accuracy increases, but for very high levels of accuracy, overconfidence increases again. More generally, the random support model predicts a nonlinear relationship between average judged probability and overall accuracy (shown in Fig. 3) that is not predicted by several other accounts of the hard–easy effect (e.g., Björkman, 1992; Suantak, Bolger, & Ferrell, 1996). This nonlinear relationship is observed in the data presented here, and is also evident in the results of other calibration research (e.g., Fig. 1 in Juslin & Olsson, 1997).

Different values for the model parameters can allow for either aggregate overconfidence or aggregate underconfidence. Consider the curve for the value $\sigma = 1$. The mean value of overconfidence (averaging over all levels of accuracy) is .063; for the mean value of overconfidence to be 0, σ needs to be about .68. This suggests that if items uniformly spanning a wide range of difficulty (from 50% to 100% accuracy) are judged, and the value of σ is constant across these items, aggregate overconfidence will be the typical result for $\sigma > .68$, and aggregate underconfidence the result for $\sigma < .68$. The random support approach can thus accommodate both overconfidence and underconfidence, with σ indexing the degree of over/underconfidence.

Relating model parameters to features of the judgment items. An additional feature of the random support approach is the ability to relate aspects of the support distributions to other features of the judgment items. For example, the estimates of δ

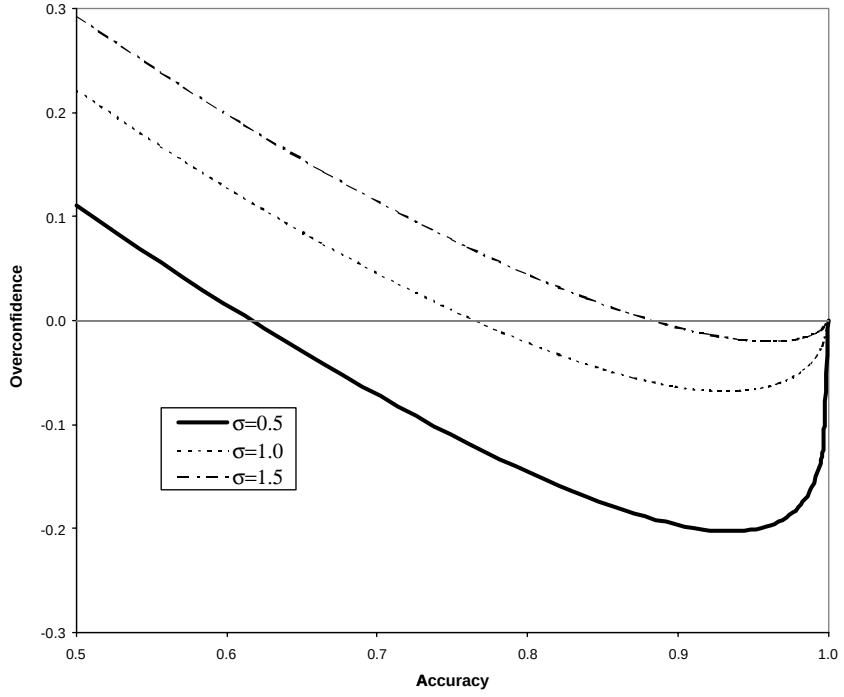

Fig. 4. Random support model predictions of the hard-easy effect: overconfidence plotted against accuracy for different values of σ.

across judged pairs of states can be related to properties of the individual states. In choosing the state with the larger population, discriminability between correct and incorrect answers (as measured by δ) would be expected to be positively related to the actual difference in population between the two states. Indeed, there is a substantial association between the log-ratio of the two states' populations and the ability of judges to discriminate between the correct and incorrect answers ($r = .78$).

Direct assessments of support. In contrast to the "psychophysical"' relationship between actual state populations and discriminability, we can also examine the correspondence between independent *subjective judgments* of individual state populations and measures of pairwise state discriminability. A separate group of 146 Stanford undergraduates rated the populations of states on a 0–100 scale. They were encouraged to use a ratio scale—to assign the most populous state a rating of 100, and then assign ratings to the other states relative to the largest state. If they believe that a state has half the population of the most populous state, its rating should be 50.

The mean values from these ratings can be considered aggregate support estimates for each individual state. Across the 25 state pairs, the association between δ and the log-ratio of the state support ratings is somewhat stronger ($r = .87$) than the correlation based on the log-true-population-ratio from the previous section ($r = .78$). Thus, independent judgments of support for single states are strongly related to the estimated separation in support distributions derived from probability judgments involving pairs of states. See Koehler (1996), Koehler et al. (1997), and Fox (1999) for additional explorations of the relationship between judged probabilities and independent assessments of support.

In short, the discriminability parameter can be predicted from the true population values of the states, or, even better, from independent judgments of state populations. Further, the discriminability parameter is more strongly related to both of these predictors than is the simple measure of percent correct ($r = .74$ between percent correct and log-support ratio; $r = .71$ between percent correct and log-population ratio). These higher correlations suggest that the discriminability parameter δ may be a better predictive measure of pairwise state judgment performance, compared to the more traditional measure of percent correct.

4. Study 2: Perceptual, cognitive, and prediction data

To explore further the generality of the random support model, and to illustrate the interpretations of the model's parameters, I fit the model to a selection of data sets from previous studies of calibration, involving several different judgment domains.

Keren (1988) studied calibration in several perceptual and knowledge tasks. In these tasks, the difficulty of the perceptual tasks was varied as follows.

- In Experiment 1, participants completed a perceptual task in which they judged whether a gap in a visually presented ring was on the right or left. This task included more difficult *small-gap* trials and easier *large-gap* trials. Participants also answered *general knowledge* questions about European geography.
- In Experiment 2, participants viewed two letters presented briefly, followed by a mask and an arrow pointing to one of the letter locations. The task was to identify which of two target letters (A or E) had been presented in the indicated location. There were three stimulus conditions. Letters were either *repeated* (the two presented letters were identical, either AA or EE), *conflicting* (both A and E were presented), or *neutral* (the uncued letter was always either K or N). The *repeated-letter inferiority effect* (Bjork & Murray, 1977; Egeth & Santee, 1981) suggests that performance will be poorer for repeated letters.

Table 1
Summary of parameter estimates and fit measures for datasets in Study 2

Data set	Condition(s)	Avg. judged prob.	$Pr(C)$	δ	σ	MAD of calib. curve	MAD of resp. prop.
Keren (1988)	Expt. 1: general knowledge	.76	.70	.74	1.08	.020	.044
Keren (1988)	Expt. 1: large gap (easy)	.68	.79	1.12	.60	.050	.048
Keren (1988)	Expt. 1: small gap (hard)	.65	.67	.61	.58	.036	.037
Keren (1988)	Expt. 2: conflict (easy)	.74	.80	1.21	.80	.012	.031
Keren (1988)	Expt. 2: neutral (medium)	.73	.77	1.05	.80	.014	.032
Keren (1988)	Expt. 2: repeated (hard)	.70	.61	.38	.83	.060	.028
Keren (1988)	Expt. 3: long exp. (easy)	.74	.77	1.04	.85	.040	.008
Keren (1988)	Expt 3: short Exp. (hard)	.65	.63	.48	.60	.028	.034
Ronis and Yates (1987)	General knowledge	.78	.67	.63	1.24	.022	.046
Ronis and Yates (1987)	Basketball	.66	.60	.34	.66	.041	.022
Wright and Wisudha (1982)	General knowledge	.77	.60	.36	1.25	.015	.059
Wright and Wisudha (1982)	Future	.77	.83	1.31	.88	.054	.038

- Experiment 3 was similar to Experiment 2, except that only the neutral letter condition was used. The exposure time of the letters was varied, to adjust the difficulty of the task from easier (*long-exposure*) to harder (*short-exposure*) while holding other aspects of the task constant.

Ronis and Yates (1987) studied calibration for general knowledge items and predictions about the outcomes of upcoming basketball games.

Wright and Wisudha (1982) studied calibration for general knowledge items and predictions about future events.

Table 1 contains the (aggregate) summary statistics, parameter estimates and fit indices for the lognormal-support model fit to each of these sets of data. The columns on the far right refer to the (weighted) mean absolute deviation between the modeled and observed calibration probabilities (''MAD of calib. curve''), and the (weighted) mean absolute deviation between the modeled and observed response proportions (''MAD of resp. prop.''). The fit between the model and data for these studies is generally very good; the average absolute deviation between model and data calibration curves and between model and data response proportions is, again, typically only about .03.

4.1. Comparison of parameter estimates

Fitting the model to these datasets illustrates how the model parameters can measure changing properties of calibration data across conditions or task types. For example, the model parameter σ provides an assessment of probability judgment extremity, controlling for accuracy. For instance, note that estimates of σ for general knowledge tasks (range: 1.08–1.25) tend to be greater than estimates of σ for

prediction (range: .66–.88) and perceptual tasks (range: .58–.85). This suggests that, for a given level of discriminability (i.e., δ, or overall accuracy), probability judgments are more extreme in the general knowledge tasks than in the prediction or perceptual tasks. This pattern is consistent with Dawes's (1980) hypothesis that overconfidence is greater for intellectual judgments than for perceptual judgments. Fischhoff and Macgregor (1982) and Wright and Wisudha (1982) noted that forecasts of future events were made with less extreme confidence than were judgments for general knowledge questions. The observation of larger values of σ for general knowledge questions than for prediction tasks is consistent with these results as well.

Furthermore, estimates of σ appear stable for different difficulty levels of similar tasks. For instance, the two perceptual tasks in Keren's Experiment 1, although differing substantially in difficulty ($\delta = 1.12$ and .61, respectively for the large-gap and small-gap judgments) yielded nearly identical values of σ, .60 and .58. Similarly, estimates of σ were very stable across the varying difficulty levels of the task in Keren's Experiment 2. However, Keren's participants appeared to be substantially more confident in their long-exposure judgments ($\sigma = .85$) than in their short-exposure judgments ($\sigma = .60$), even after adjusting for the varying difficulty of the two types of judgments. One interpretation is that the short-exposure judgments seemed much more difficult to the participants (even adjusting for the actual difficulty), prompting more conservative ratings of confidence, and thus a lower value of σ.

In summary, the random support model provides a good fit to perceptual judgments, cognitive judgments, and predictions of future events. Analyses of the model parameter estimates tentatively suggest: (a) consistent judgment extremity across related tasks of varying difficulty, and (b) lower judgment extremity for future predictions and perceptual judgments, compared to higher judgment extremity for general knowledge judgments.

5. Study 3: Three-alternative state judgments

As a further extension, I now explore the generalizability of the model to three-alternative forced choice (3AFC) tasks. In this study, participants judged the populations of either pairs or triples of states. Based on the parameter estimates obtained from fitting the random support model to the state-pair judgments, predictions for the state-triple judgments are derived and compared to the data. Thus, like the analysis of independent judgments of support from Study 1, this study allows an out-of-sample test of the model's fit; parameters are estimated from one set of data, and predictions are evaluated based on an independent set of data from a different task.

5.1. Method

Participants. Participants were 121 Stanford University undergraduates enrolled in an introductory psychology course. They participated in the study for course credit, completing the questionnaire in a packet composed of several other unrelated tasks.

Design. One group of participants (group T, $n = 65$) was presented with a set of US state triples. For example, one triple consisted of Georgia, Nevada, and Kansas. For each triple, participants were instructed to pick the state with the largest population and then rate the probability that their chosen answer was correct, using a scale ranging from 33% to 100%.

Two other groups of participants (P1 and P2, $n = 27$ and 29, respectively) completed a set of similar judgments involving *pairs* of states. For each item for Groups P1 and P2, the correct answer from each triple in Group T was paired with one of the incorrect answers from the triple. For example, the state pairs corresponding to the

triple described above would be (Georgia, Nevada) and (Georgia, Kansas). Each participant made judgments for either 25 state pairs or triples.

5.2. Results

First, overall accuracy rates and average judged probabilities for each of the 25 state triples are compared to the predictions of the random support model. Second, the overall calibration curve across all triples is compared with the model's predictions.

Item-level accuracy. The random support model was first fit to the data from groups P1 and P2. For example, estimates of δ and σ were derived from data for the Georgia–Kansas, and, separately, from data for the Georgia–Nevada pair. In this manner, the differences in means between the support distributions for Georgia and Kansas, and for Georgia and Nevada, can be estimated. Because the correct answer (e.g., Georgia) was present in both pairs, the means of the three support distributions can be scaled based on the two δ estimates from groups P1 and P2. An average estimate of σ was computed from the values derived from groups P1 and P2. Using these three estimated parameters, the overall accuracy rate and average judged probability for each state triple can be predicted, and compared to the corresponding observed judgments.

In Fig. 5, the observed accuracy for each state triple is plotted against the item-level accuracy predictions of the model. For comparison, also displayed are the fits

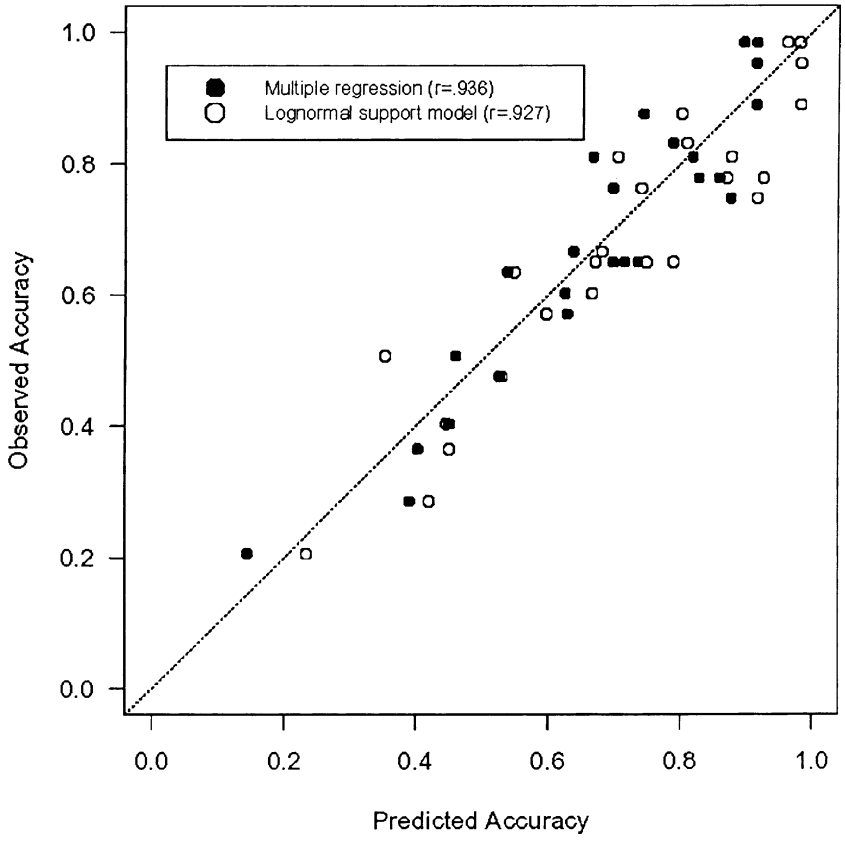

Fig. 5. Observed vs. predicted accuracy rates for state triples in Study 3. Predictions for "lognormal support model" (open circles) are based on the random support model using parameters derived from the two state pair conditions. "Multiple regression" predictions (filled circles) are derived from regressing triple accuracy rates on the two state pair accuracy rates.

based on a multiple regression of the triple-accuracy rate predicted from the two pairwise accuracy rates (from groups P1 and P2). The correlation between the random support model predictions and the actual triple accuracy measures is very high ($r = .93$) and essentially the same as the multiple correlation of the best linear fit based on the two pairwise accuracy measures ($R = .94$), even though the latter model requires the estimation of three additional parameters.

Item-level mean judged probability. In Fig. 6, the observed average judged probability for the triple items is plotted against the predictions of the model. Also plotted, for comparison, is the multiple regression fit of the average judged probability of the triples predicted from the average judged probability of the two corresponding pairs. Again, the predictions of the model are excellent ($r = .95$) and correlate virtually as well as do the multiple regression fits, which involve fitting three additional parameters ($R = .96$). The mean absolute deviation between the model predictions of average judged probability and the data is a mere .027.

Overall calibration. Finally, the pairwise judgment data were used to generate a predicted calibration curve for the triples data. The item-level parameter estimates obtained from the pairs were used to generate the model's predictions for the triples judgments. Because the distribution of probability judgments (or log-odds) for the triples case is not easily related analytically to the normal distribution, a simulation approach was used instead. For each of the state triples, 10,000 probability judgments were simulated based on the δ and σ parameters estimated from the pairs data. Fig. 7 displays the aggregate calibration curve based on these simulations, as well as

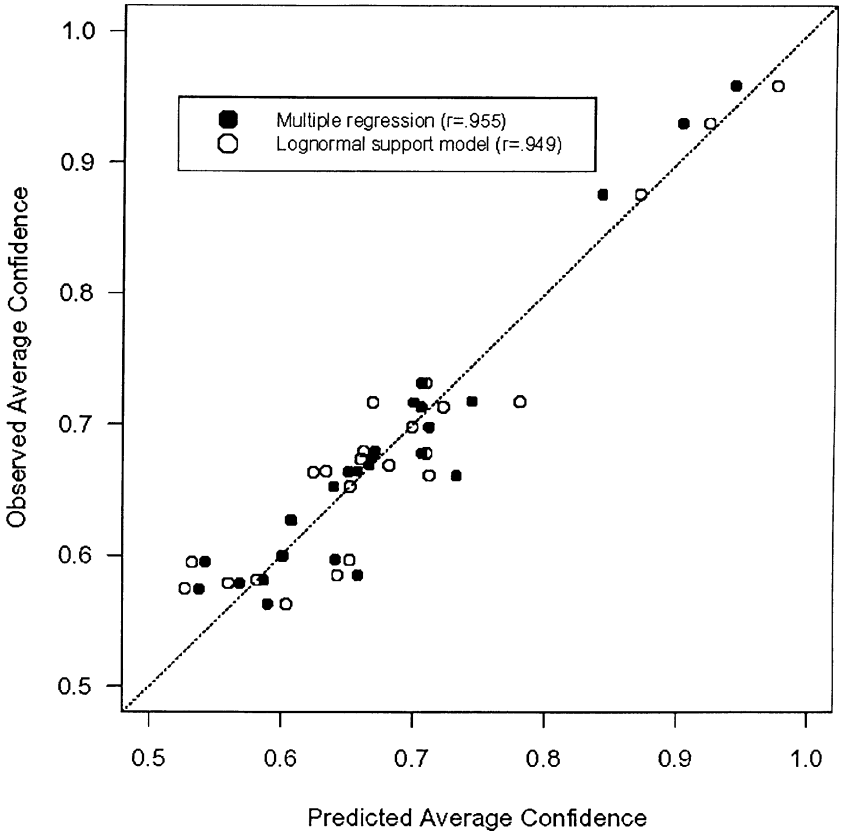

Fig. 6. Observed vs. predicted average confidence for state triples in Study 3. Predictions for "lognormal support model" (open circles) are based on the random support model using parameters derived from the two state pair conditions. "Multiple regression" predictions (filled circles) are derived from regressing average confidence for the triples on the two state pair average confidence scores.

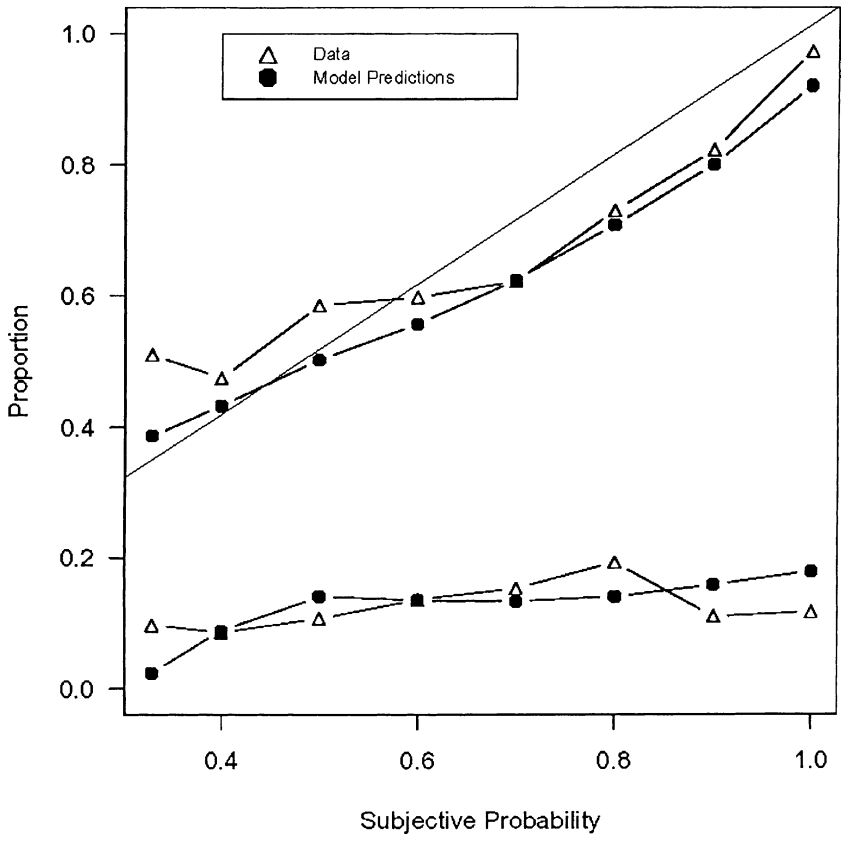

Fig. 7. Data (triangles) and predicted (filled circles) calibration curves and response proportions for state triples in Study 3. All model predictions are based on parameters derived from the state pairs data.

the actual data from triple judgments. Again, the model fits quite well; average absolute deviation between predicted and observed calibration curves is .040 and between predicted and observed response proportions is .036. Note that in all the tests involving triple judgments, unlike tests for the 2AFC tasks considered previously, the model is predicting completely new data and no new parameters are being estimated.

6. Comparison of models of calibration

As noted earlier, the random support model is similar in spirit to several other stochastic models of calibration. Below, I compare and contrast some of the features of these models.

6.1. Decision variable partition model

Ferrell and McGoey (1980) proposed an stochastic "signal detection" model of the calibration of probability judgments which is closely related to the random support model. For a two alternative forced choice (2AFC) task, for instance, their model proposes that the judge experiences an internal measure of "subjective certainty" (Ferrell & McGoey, 1980, p. 33) for each option. As in the present random support model, the option which produces greater subjective certainty is chosen as the answer. The judged probability that the answer is correct is based on the absolute value of the difference between the two subjective certainty measures. This decision

variable is partitioned into k-ordered confidence ratings (e.g., .50, .60, etc.) by a set of $k-1$ cutoffs. In this model, the confidence categories comprise an ordinal scale, and do not necessarily refer to numerical probability judgments; they could refer, for example, to ordered verbal probabilistic labels (e.g., unlikely, possible, likely).

In Ferrell and McGoey's decision variable partition (DVP) model, the feelings of subjective certainty for correct and incorrect answers are modeled as random variables from normal distributions with unit variance. The separation between the two distributions of subjective certainty is the measure of discrimination—the ability to tell the difference between correct and incorrect answers. In the 2AFC case, the parameters of the DVP model are the measure of discriminability and the set of cutoffs which partition the decision variable into classes of probability judgments. In the case of 6 confidence levels, 6 parameters are used to estimate 11 probabilities (6 conditional "calibration" probabilities and 5 unconstrained response proportions). Ferrell and his colleagues have successfully used the model to describe changes in calibration due to varying item difficulty or base rate (Smith & Ferrell, 1983), and due to performance feedback and training (Ferrell, 1994b; Ferrell & McGoey, 1980). Ferrell has discussed how his model can be successfully applied to both cognitive and sensory/perceptual tasks (Ferrell, 1995, 1994a; Suantak et al., 1996).

The random support model is similar to Ferrell's model in the following respects:

1. Subjective certainty for each option is modeled as a random variable, with subjective certainty for correct options typically greater than subjective certainty for incorrect options.
2. The option with the greater subjective certainty is chosen by the subject.
3. The subjective certainty values for the two options are combined to produce a probability judgment.

However, the present model also differs from Ferrell's model in several important ways.

1. Subjective certainty is interpreted as support for the considered hypotheses, in the context of support theory.
2. Judged probability is mapped directly from the support values for the two options, rather than using multiple cutoffs.
3. Only two parameters are required to model the calibration curve and distribution of response proportions for the 2AFC case. No variable cutoff parameters are used. In effect, the "cutoffs" are implicitly instantiated in the support distributions and the transformation from support into judged probability.
4. The random support model is easily extended to judgments between more than two alternatives.
5. Parameters of the support distributions can be linked naturally to other features of the judgment items or characteristics of the judge(s).

In these ways, the random support model can provide a more parsimonious account of numerical probability judgment than does Ferrell's model. However, Ferrell's model has the advantages that (a) ordinal judgments of likelihood can be accommodated easily, and (b) the additional cutoff parameters allow for greater flexibility in fitting data.

Griffin and Tversky (1992) proposed a model quite similar to Ferrell's model (as noted by McClelland & Bolger, 1994), except incorporating discrete rather than continuous underlying random variables. In Griffin and Tversky's model, for a particular two-alternative question, the subject searches memory and samples units of "evidence", as if drawing from an urn containing a mixture of red and white balls. Whichever option receives more units of evidence (red balls) is chosen as the answer, and the reported probability judgment is the proportion of total evidence in favor of the chosen option. The number of units of evidence are modeled as binomial random variables. Griffin and Tversky introduce this model in the context of their discussion of the strength and weight of evidence. Strength refers to the proportion of the

sample of evidence favoring one option over another, and weight refers to the size of the sample drawn. In this simple chance model, the reported probability judgment only reflects the strength of the evidence considered (i.e., the proportion of evidence favoring the chosen option) and does not reflect the weight, or credence, of the evidence. Griffin and Tversky show that their simple model of judgment, incorporating a varying discriminability parameter, can describe distinct qualitative patterns of calibration data across different task characteristics. Following Griffin and Tversky's premise of confidence judgments driven primarily by strength rather than weight, Koehler, Brenner, and Griffin (2002) apply the present random support model to patterns of calibration by experts in domains such as medicine, sports, and business.

Pfeifer (1994) also proposed a discrete-variable stochastic account of calibration. He assumes that for a particular question, the probability of answering correctly is p. The subject reviews the total set of knowledge she has about the question (n) and notes the amount of evidence (r) favoring the focal event. The reported probability judgment is then $\hat{p} = r/n$. For instance, a weather forecaster assessing the probability of rain might implicitly consider (from memory) the n days sharing similar conditions to the present day, and recall rain on r of those days. The evidence for the focal event need not be a simple tally of frequencies as in this example, but rather evidence more generally defined.

Pfeifer considers the case where judges are unbiased, in the sense that $E(\hat{p}) = p$, and r follows a Binomial (n, p) distribution. In this case, he argues, the observed calibration data may still suggest overconfidence, as a consequence (or illustration) of regression toward the mean.

6.2. Regression-to-the-mean "error" models

Pfeifer's conclusion is similar in this respect to the arguments of Erev et al. (1994) (also Budescu, Erev, & Wallsten, 1997a; Budescu, Wallsten, & Au, 1997b). They show that if subjects' probability judgments $\hat{p}$ are centered around the "true" probability p but are perturbed by error, calibration data may appear to show spurious overconfidence.

These accounts assume that subjects in some sense have internalized the "true" probabilities they are trying to estimate. However, the translation to an expressed probability judgment may introduce error. The random support account makes no such assumption about access to the true probabilities, or underlying unbiased judgment. Rather, in the random support model, subjects simply weigh evidence for the propositions under consideration, and report a judgment based on this evidence. The random support approach does include a model of variability of judgment, but the variability is embodied initially in the distributions of support, rather than judged probability.

One common conclusion taken away from the error models is that overconfidence in stated probabilities may partially be a statistical artifact, a consequence only of error in the translation from covert, internal confidence to an overt, stated probability. A common way of stating this argument is to say that in some cases overconfidence may not be "real." This conclusion may be misleading in that it appears to suggest that, given that in the past when a judge makes predictions with 90% confidence and only 70% of the predictions turn out to be correct, one should nonetheless expect future predictions to be well-calibrated. This interpretation would be mistaken, unless the "error" inherent in the judge's stated probabilities somehow disappeared. More generally, it is unclear why one should evaluate the quality of calibration on the unobservable, latent "true scores" that are constructed in these error models, and allow those true scores to trump the observed probabilities as the quantities to be evaluated for good calibration. A more detailed critique of this

interpretation derived from error models can be found in Brenner (2000); see also the reply by Wallsten, Erev, and Budescu (2000).

6.3. Stochastic judgment model

Budescu et al. (1997b), following Wallsten and Gonzalez-Vallejo's (1994) stochastic model of statement verification, propose a model of probability judgment which is in many ways similar to the random support model and the decision variable partition model of Ferrell and McGoey (1980). Like the decision variable partition model, their stochastic judgment model (SJM) focuses primarily on the mapping from covert feelings of confidence to overt responses via variable response cutoffs. The random support model differs from this model primarily by avoiding the use of multiple cutoffs, and incorporating a direct mapping from support (i.e., covert confidence) to judged probability. As a result, the random support model can be easily applied to tasks involving three or more propositions, while extending the stochastic judgment model to such tasks is less straightforward.

6.4. Ecological models

Several authors (Björkman, 1994; Gigerenzer et al., 1991; Juslin, 1994) have proposed what have been termed "ecological" models (McClelland & Bolger, 1994) of the calibration of probability judgments, citing the work of Brunswik (1943, 1955). The main premise of these models is that people internalize the associations between cues and events in the world (variously termed ecological validities or external cue validities), and draw upon this internalized knowledge when judging likelihoods. With exposure over time to a particular judgment domain, the internal cue validities are assumed to closely match the ecological validities.

The ecological models posit that if people appropriately internalize ecological validities (and the proponents of these models generally assume that they do), people should be well-calibrated on judgment items that are representative of the overall judgment domain. If, however, the judgment items are selected by experimenters to be especially difficult, or are particularly surprising or otherwise non-representative of a natural reference class to which people have adapted, people's internal cue validities will generally overestimate the predictive validity of the cues in the sample of judgment items, and overconfidence will result. In short, the ecological models attribute overconfidence to non-representative selection of judgment items or tasks.

In the context of the random support model, surprising or misleading items are those for which δ is negative. More generally, it is easy to represent "biased" sets of judgment items for which predictive cues are less predictive than they are in the full population of items. We could imagine distributions representing support for all correct and all incorrect answers across a relevant population of potential judgment items (e.g., pairs of cities or states). Call these the *population support distributions*. If the correct and incorrect support distributions for a particular sample of judgment items are closer together than the population support distributions, this is analogous to what the ecological proponents consider a biased or non-representative set of items. Using different support distributions, the random support approach can represent the match or mismatch between cue validities for a set of items and the corresponding ecological validities in the entire population of items.

But in contrast to the ecological models, the random support approach does not assume that if items are randomly sampled from an appropriate reference class, overconfidence should disappear. While the notion of biased sampling of items is clearly important, the claim that random sampling of items eliminates overconfidence has not always been supported (Brenner, Koehler, Liberman, & Tversky, 1996; Griffin & Tversky, 1992). The random support approach can represent biased

or representative item selection without requiring that overconfidence disappear in the case of the latter.

Gigerenzer et al. (1991) and Juslin et al. (2000) also argue that the hard-easy effect should disappear given random selection of judgment items at different levels of difficulty. The random support model, in contrast, predicts the hard–easy effect for most of the range of item difficulty, regardless of random selection of items. Data from Study 1 and other sources (e.g., McClelland & Bolger, 1994) document the existence of the hard–easy effect even when items are randomly sampled from a relevant reference class.

7. Summary and conclusions

The proposed random support model of probability judgment was shown to account for a wide array of calibration data. The model provides a unified account of calibration data from tasks with varying numbers of alternatives. Independent judgments of support correlate with parameters derived from probability judgments, validating the construct of support. Parameters of the model are easily interpretable and permit comparisons of different aspects of the quality of probability judgments across judge and task characteristics. The random support model differs from previous stochastic models of calibration by eliminating the use of variable cutoff parameters, mapping underlying support directly into judged probability, and by providing a very general framework (support theory) for the analysis of different judgment tasks.

Acknowledgments

I am indebted to Gordon Bower, Dale Griffin, Derek Koehler, Lee Ross, Seenu Srinivasan, Tom Wallsten, and especially Yuval Rottenstreich, Ewart Thomas, and the late Amos Tversky for many helpful comments on the work presented in this article. Portions of this paper are based on a dissertation submitted in partial fulfillment of the Ph.D. requirements at Stanford University.

References

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., Wallsten, T. S., & Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied, 6*, 130–147.

Bjork, E. L., & Murray, J. T. (1977). On the nature of input channels in visual processing. *Psychological Review, 84*, 472–484.

Björkman, M. (1992). Knowledge, calibration, and resolution: A linear model. *Organizational Behavior and Human Decision Processes, 51*, 1–21.

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. (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., & 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., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes, 65*, 212–219.

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*. New York: Cambridge University Press.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review, 50*, 255–272.

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

Budescu, D. V., Erev, I., & Wallsten, T. S. (1997a). On the importance of random error in the study of probability judgment. Part I: New theoretical developments. *Journal of Behavioral Decision Making, 10*, 157–171.

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997b). 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.

Dawes, R. (1980). Confidence in intellectual judgments vs. confidence in perceptual judgments. In E. D. Lantermann, & H. Feger (Eds.), *Similarity and choice: Papers in honor of Clyde Coombs* (pp. 327–345). Bern: Hans Huber Publishers.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory process model for judgments of likelihood. *Psychological Review, 106*, 180–209.

Egeth, H. E., & Santee, J. L. (1981). Conceptual and perceptual components of interletter inhibition. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 506–517.

Erev, L., 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. (1994a). Calibration of sensory and cognitive judgments: A single model for both. *Scandinavian Journal of Psychology, 35*, 297–314.

Ferrell, W. R. (1994b). Discrete subjective probabilities and decision analysis: Elicitation, calibration and combination. In G. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 411–451). Chichester: John Wiley and Sons.

Ferrell, W. R. (1995). A model for realism of confidence judgments: Implications for underconfidence in sensory discrimination. *Perception and Psychophysics, 57*, 246–254.

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

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

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

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

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

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.

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

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

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., 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*. New York: 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.

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

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980–1993. In G. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). Chichester: John Wiley and Sons.

Pfeifer, P. E. (1994). Are we overconfident in the belief that probability forecasters are overconfident? *Organizational Behavior and Human Decision Processes, 58*, 203–213.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes, 40*, 193–218.

Rottenstreich, Y., Brenner, L., & Sood, S. (1999). Similarity between hypotheses and evidence. *Cognitive Psychology, 38*, 110–128.

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

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.

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

Spetzler, C. S., & Staël von Holstein, C.-A. S. (1975). Probability encoding in decision analysis. *Management Science, 22*, 340–358.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard–easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes, 67*, 201–221.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286.

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

Wallsten, T. S. (1996). An analysis of judgment research analyses. *Organizational Behavior and Human Decision Processes, 65*, 220–226.

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

Wallsten, T. S., Erev, I., & Budescu, D. V. (2000). The importance of theory: Response to Brenner (2000). *Psychological Review, 107*, 947–949.

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

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

Yates, J. F. (1994). Subjective probability accuracy analysis. In G. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 381–410). Chichester: John Wiley and Sons.

Received 27 July 1999